School Effectiveness and School Improvement
An International Journal of Research, Policy and Practice

ISSN: 0924-3453 (Print) 1744-5124 (Online) Journal homepage: http://www.tandfonline.com/loi/nses20

Diversity in school performance feedback systems

Goedele Verhaeghe, Kim Schildkamp, Hans Luyten & Martin Valcke

To cite this article: Goedele Verhaeghe, Kim Schildkamp, Hans Luyten & Martin Valcke (2015) Diversity in school performance feedback systems, School Effectiveness and School Improvement, 26:4, 612-638, DOI: 10.1080/09243453.2015.1017506

To link to this article: http://dx.doi.org/10.1080/09243453.2015.1017506

Published online: 26 Mar 2015.

Submit your article to this journal

Article views: 360

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=nses20

Download by: [Universiteit Twente] **Date:** 11 November 2015, At: 02:33

School Effectiveness and School Improvement, 2015
Vol. 26, No. 4, 612–638, http://dx.doi.org/10.1080/09243453.2015.1017506

Routledge
Taylor & Francis Group

Diversity in school performance feedback systems

Goedele Verhaeghe[a], Kim Schildkamp[b]*, Hans Luyten[c] and Martin Valcke[d]

[a]*Ministry of Education, Brussel, Belgium;* [b]*ELAN, University of Twente, Enschede, the Netherlands;* [c]*Department of Research Methodology, Measurement and Data Analysis, University of Twente, Enschede, the Netherlands;* [d]*Department of Educational Studies, Ghent University, Ghent, Belgium*

(*Received 21 August 2013; final version received 5 February 2015*)

As data-based decision making is receiving increased attention in education, more and more school performance feedback systems (SPFSs) are being developed and used worldwide. These systems provide schools with data on their functioning. However, little research is available on the characteristics of the different SPFSs. Therefore, this study reflects on the characteristics of SPFSs to provide feedback designers and users arguments for making sound choices in selecting SPFSs with particular characteristics. The results of our study show that the 5 SPFSs selected for the purpose of comparison differ with respect to features related to data gathering and data analysis processes, the content, and the numerical measures and representation modes used. A wide variety can be detected in terms of the complexity and accuracy of data modeling. Users need to be properly informed about the underlying rationale for the features of each SPFS, and on the limitations and strengths of the performance indicators used.

Keywords: data-based decision making; data-driven decision making; data systems; school performance feedback systems; data analysis

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

Introduction

In most countries around the world, schools are required to systematically gather data about their school functioning, comprising elements such as educational process and performance outcomes. Data can be defined as quantitative as well as qualitative information that is systematically collected and organized to represent some aspect of schooling (Lai & Schildkamp, 2012; Wayman, Jimerson, & Cho, 2012). Examples of these data are assessment data, structured classroom observation data, and student background information. These data provide a school with performance feedback, which can be used in decision-making processes. When decisions are based on data, this is called data-based decision making (in short, data use) (Lai & Schildkamp, 2013; Mandinach & Honey, 2008).

Data use is a complex and interpretive process, involving data gathering, data analyses and interpretation, and taking action based on data (Coburn, Toure, & Yamashita, 2009; Coburn & Turner, 2012). The action's impact should be evaluated by gathering new data, which creates a feedback loop (Mandinach & Jackson, 2012). The quality of the actions taken based on the data is, among other things, dependent on the quality of the data (Coburn & Turner, 2011; Schildkamp & Kuiper, 2010). Therefore, access to high-quality

*Corresponding author. Email: k.schildkamp@utwente.nl

© 2015 Taylor & Francis

data is essential (Breiter & Light, 2006; Wayman & Stringfield, 2006). These data can be gathered with school performance feedback systems (SPFSs).

School performance feedback systems

Many schools, in countries all over the world, use school performance feedback systems (SPFSs) to gather these data, which "are information systems external to schools that provide them with confidential information on their performance and functioning as a basis for school self-evaluation" (Visscher & Coe, 2002, p. xi). The following aspects are important characteristics of SPFSs (Visscher & Coe, 2002):

- The systemic organization of the feedback initiative: The feedback providers are bound to an organization and produce school performance feedback not as a one-shot activity but on a systematic basis.
- The external component: This refers mainly to the data analysis and feedback provision. The data gathering process can be conducted in cooperation with school team members.
- The goal of school improvement: This implies that SPFS developers provide the school performance feedback on a confidential basis, in contrast with information made public for accountability reasons. By generating data for voluntary use by schools, SPFSs are considered as professional monitoring systems. They differ from official accountability systems, by which schools are hold accountable as publicly funded institutions (Tymms, 1999).
- The unit level of information: School performance feedback goes beyond individual pupil results. At least some indications are provided on the schools' functioning and effectiveness by aggregating data.
- The content of the feedback: The content refers to the schools' performance and functioning. A schools' functioning encompasses more than merely output results, but also refers to context-, input-, and process-related indicators.

If one looks at the definition and characteristics of a SPFS, many different systems might be considered as SPFSs, including central examination systems, school inspectorate, national assessment systems, pupil monitoring systems, research projects, school self-evaluation systems, and providers of standardized tests (see Table 1).

Though all systems described in Table 1 can function as SPFSs, they simultaneously might also function as an official accountability system. For example, central examination data are often considered by inspectorates and parents as a performance indicator for the school's functioning. In addition, these data can be transformed into confidential feedback, after having performed secondary analyses on these results (Yang, Goldstein, Rath, & Hill, 1999). Also, reports from a school inspection visit can serve both purposes of accountability and improvement. This illustrates that the relation between accountability and improvement may have different configurations (Earl & Fullan, 2003; Hofman, Dijkstra, & Hofman, 2009; Maier, 2010; Vanhoof & Van Petegem, 2007; Zupanc, Urank, & Bren, 2009). A tension always exists between using data for accountability and improvement purposes. Using data for accountability is often complicated by pressures to perform, test pollution, and punitive actions such as "naming, shaming, and blaming", which subvert the purpose of the system and prevents data use for improvement purposes (Archer, 2010; Hattie et al., 2005; Jansen, 2001; Taylor, 2007). This paper therefore focuses on the use of data from SPFSs for improvement purposes.

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

Table 1. Different kinds of SPFSs.

SPFS	Description
School feedback projects	The core task of these systems is providing schools with confidential information on their functioning.
Central examination systems	Sometimes (raw or adjusted) results of central examinations are fed back to schools for school improvement, instead of/in addition to making the results public.
School inspectorates	These reports can be considered as school feedback if they serve the purpose of school improvement, in addition to accountability.
National assessment systems	This differs from central examinations as this information is gathered for by governments to put measures of school performance into the public domain at a national educational level. However, if school-specific results are confidentially fed back to schools, it can be considered as school feedback.
Pupil monitoring systems	These systems are developed to assess individual pupils'/students' learning progress. These results can be used as school feedback, when also aggregated reports are provided for a group of pupils/students.
Research projects	Participation in research projects can result in a school feedback report, as a return in investment.
School self-evaluation systems	These are systems developed only with the purpose to provide schools with confidential information on their performance and functioning.
Providers of standardized tests	Some (psychometric) standardized tests, taken from individual pupils/students, can result in aggregated scores for a class or group and thus can be considered as school feedback.

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

SPFSs primarily aim at supporting school improvement. These feedback initiatives contribute to the creation of information-rich environments, which are essential for schools in their data-based decision making. Although data from SPFSs are only one source of information, they provide schools with important data on variables associated with school effectiveness, which schools can use to improve their performance in terms of improving teaching and ultimately student achievement (Davies & Rudd, 2001; Visscher & Coe, 2003).

However, although the use of data can lead to increased student achievement (Campbell & Levin, 2009; Carlson, Borman, & Robinson, 2011; Lai, McNaughton, Timperley, & Hsiao, 2009), the empirical findings do not always confirm the expected positive effects of SPFSs. Several studies (Schildkamp & Kuiper, 2010; Schildkamp & Teddlie, 2008) show that often the actual within-school use of school performance feedback remains limited, which may (partly) be caused by the characteristics of these SPFSs or lack of a good functioning SPFS (Breiter & Light, 2006; Chen, Heritage, & Lee, 2005; Coe & Visscher, 2002; Datnow, Park, & Wohlstetter, 2007; Earl & Fullan, 2003; Kerr, Marsh, Ikemoto, Darilek, & Barney, 2006; Schildkamp & Kuiper, 2010; Schildkamp & Visscher, 2009; Sharkey & Murnane, 2006; Verhaeghe, Vanhoof, Valcke, & Van Petegem, 2010; Wayman, 2005; Wayman, Cho, & Johnston, 2007; Wayman & Stringfield, 2006; Wohlstetter, Datnow, & Park, 2008).

All SPFSs adopt their own data gathering systems, statistical methods, data representations, and so forth. However, little is known about the distinct characteristics of these SPFSs, or of the rationale behind these features. Little is also known about whether its users are capable of correctly interpreting and analyzing data derived from these systems, which is a crucial condition for data-based decision making for improvement purposes, as well as for accountability purposes. A debate on characteristics of SPFSs could be a starting point for reflection for current and future feedback providers and users. Therefore, this study has been set up, focusing on comparing the characteristics of different SPFSs. We will examine a number of diverse SPFSs and the underlying rationale for these variations between systems.

The central aim of this study is to explore the diversity in (technical) characteristics of SPFSs and to reveal the underlying rationale. Literature on SPFSs reveals that the analytical framework developed by Visscher (Visscher, 2002; Visscher & Coe, 2003) is the most frequently cited and used (e.g., in Hellrung & Hartig, 2013; Maier, 2010; Schildkamp & Teddlie, 2008; Schildkamp & Visscher, 2009; Verhaeghe et al., 2010; Zupanc et al., 2010). This framework discerns a set of factors influencing the use of the performance feedback, including the design features of the underlying SPFSs and the characteristics of the feedback report itself. We used this framework as a basis, focusing on the technical aspects of SPFSs (see Table 2). These technical aspects, for example, data gathering, data analyses, subsequent content of feedback, and graphical representations used, are crucial aspects of effective data use (Kerr et al., 2006: Sharkey & Murnane, 2006; Visscher, 2002; Wayman et al., 2007; Wayman & Stringfield, 2006).

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

Data gathering, data analysis, feedback content, and graphical representations

SPFSs gather data about a school's instructional process and performance, by making use of performance indicators. Following Goldstein and Spiegelhalter, "a performance indicator is a summary statistical measurement on an institution or system which is intended to be related to the 'quality' of its functioning" (1996, p. 385). Rowe and Lievesley add an evaluative component to this definition: "performance indicators (PIs) are defined as data indices of information by which the functional quality of institutions or systems may be measured and evaluated" (2002, p. 1). Applied to the context of schools, Fitz-Gibbon and Tymms (2002) define an indicator "as an item of information collected at regular intervals to track the performance of a system" (p. 2). Hereby, they emphasize the systematic character of the data gathering and analysis, which corresponds to the definition of SPFSs by Visscher and Coe (2002). School performance indicators do not only report about the output aspect of school quality, such as pupil achievement results, but also on the context, input, and process of the school's functioning. These can include indicators on resource provision and funding, participation rates of pupils, repetition rates, class sizes, factors affecting students' progress rates, and so forth (Rowe & Lievesley, 2002).

With regard to this data gathering process, it is important to look at the persons gathering the data (e.g., SPFS field workers or school team members), the types of instruments used (e.g., assessments, surveys), the data gathering medium (e.g., paper and pencil, computer-based), time and place of data collection, and the data source.

To successfully serve schools in their internal quality policy, data on these indicators need to be analyzed. The feedback resulting from these analyses has to meet certain criteria (Fitz-Gibbon, 1996; Heck, 2006; Rowe, 2004; Rowe & Lievesley, 2002; Schildkamp & Teddlie, 2008; Visscher, 2002). First, feedback needs to be relevant and useful, which means it corresponds to the actual information needs of the users.

Table 2. Comparing technical SPFS characteristics.

SPFS characteristics	Items
Data gathering	– Data administrators (e.g., school team members, field workers from SPFS) – Medium (e.g., paper pencil, computer) – Structuredness of instruments (e.g., completely structured, semi-structured, computer adaptive) – Types of instruments (e.g., tests, interviews, surveys, observation scales) – Data source (e.g., pupils, teachers, parents) – Timing (e.g., any time, fixed moments) – Place (e.g., classroom, computer lab, playground) – Options in test administration (e.g., fixed, flexible, or demand-driven supply)
Data analysis	– Type of analysis (e.g., quantitative, qualitative) – Scaling model (e.g., classical test theory, item response theory) – Model used (e.g., regression model, ordinary least squares, multi-level analysis) – Type of value added (e.g., prior, concurrent) – Levels of unit (e.g., pupil level, year group level, school level, cohort level, subscale level, item level, subject level, aggregate level) – Measurement moments (e.g., single measurements, successive measurements, two linked measurements, longitudinal measurements)
Feedback content	– Variables (e.g., attitudinal, behavioral, cognitive, contextual) – Subjects (e.g., language, mathematics, science, world orientation) – Non-subject-specific information (e.g., school culture, pupil background variables, pupil mobility, socioemotional development, ADHD scale, attitudes to school, dyslexia, study skills) – Numerical measures (e.g., raw scores, cut-off score, gain score, mean score, value-added score) – Reference group (e.g., national average, representative sample of population, group of participating schools) – Type of reference (e.g., self-referenced, norm referenced, criterion referenced) – Reliability indication (e.g., confidence intervals, significant values) – Text content (e.g., results, interpretation of results, explanation of statistical concepts and graphical representations, information on how to communicate results)
Graphical representation	– Feedback medium (e.g., static reporting, flexible tool) – Graphical representations (e.g., bar graph, box plot, histogram, layer graph, line graph, pie graph) – Reliability indices (e.g., confidence intervals, significance values)

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

Furthermore, feedback needs to be accurate, which relates to the reliability and validity of the data gathered. Related to this utility perspective, the performance indicators should be delivered timely, which both concerns the currency and punctuality of the delivered feedback. Furthermore, users need to accept the performance indicators and consider them to be fair. This fairness does not only refer to the striving towards unbiased results but also to the interpretability, reliability, stability, and incorruptibility of the reported performance indicators. Lastly, performance indicators should strive towards beneficial

effects and should avoid unwarranted harm (Fitz-Gibbon, 1996; Fitz-Gibbon & Tymms, 2002; Goldstein & Myers, 1996). Therefore, in order to develop insight about the relevance of the feedback system, the contents of the feedback reports of the selected systems are described in this study, including the feedback representations used. This includes both the numerical measures and graphical representations used, to get a view on the interpretability of what is fed back to schools.

Gathering information about these technical aspects is important with regard to the relevance, usefulness, accuracy (reliability and validity), timeliness, and perceived fairness of the data. Studies show that access to data, for example, provided by SPFSs, that are easy to analyze and entail feedback that is clear and understandable are more likely to lead to an increased level of data use. Data use is likely to be constrained if schools have difficulties in gathering the data they need, have difficulties in analyzing the results, or in understanding the feedback representations used (Breiter & Light, 2006; Chen et al., 2005; Coburn & Turner, 2011; Park & Datnow, 2009; Schildkamp & Kuiper, 2010; Wayman & Stringfield, 2006; Wohlstetter et al., 2008). Therefore, in this study we will compare different SPFSs with regard to the (a) data gathering processes these use, (b) data analyses, (c) the feedback content, and (d) the graphical representations used.

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

Method

Selected SPFSs

The five systems described in this study were purposefully selected. We selected systems that consisted of different types of (assessment data). Moreover, we selected systems that have been researched. The selection is therefore not representative, but illustrative of the variance in SPFSs adopted in educational settings. First, each selected SPFS is shortly described (more extensive details will be provided in the results section):

- **Assessment tools for teaching and learning (asTTle):** AsTTle has been developed as part of a government-funded research project at the Visible Learning Labs of the University of Auckland in New Zealand. This SPFS offers schools a national assessment model with all characteristics of a SPFS, without the consequences of high-stakes testing. The system is an electronic test creation and reporting engine and test item bank covering reading, writing, and mathematics in both English and Maori for pupils aged 4 to 17. AsTTle allows teachers to customize standardized 40-minute pencil-and-paper or computer-based tests according to their priorities for test difficulty and content, regardless of the year or age of students. The asTTle tools provides teachers and school leaders with the ability to analyze achievement of individual students or groups/subgroups of students, gain insights as to strengths and weaknesses, and points to additional teaching and curriculum resources through an online catalogue. The feedback helps teachers to get acquainted with the national curriculum, and is aimed at enhancing teaching and learning. About 80% of all elementary and high schools of NZ are using asTTle (Years 4–12). Participation is voluntary and free of charge. The feedback is offered both in English and Maori, which have two distinct curricula. Feedback reports are delivered directly and immediately to school team members and pupils/students and parents via a secured online website or via software used on the local network. Results are not made public. AsTTle offers direct feedback delivery to students and parents. The technological applications allow pupils to get access to their results during their school

career, over all different years and schools. AsTTle functions as a professional monitoring system as the purpose is to create a low-stake assessment system to be used internally within the schools. Its main function is the detection of learning needs on an aggregate level (Hattie, Brown, & Keegan, 2003; Hattie et al., 2005).

- **Performance indicators in primary schools (PIPS):** PIPS was developed by The Centre for Evaluation and Monitoring at the Durham University (UK). It provides schools with annual assessments (e.g., maths, literacy) and provides schools with measures of value added achievement for pupils aged 4 to 11. It also includes data from observation, an ADHD (Attention Deficit Hyperactivity Disorder) scale, and pupil background information. It is widespread in primary schools (from reception to Year 6) in England and Scotland, and to a smaller scale in other parts of the UK. Furthermore, PIPS has local adaptations of the system, applied worldwide. Within the UK, independent schools show the largest interest in PIPS, as compared to the government-funded schools, as they lack monitoring systems and information on national testing, because they do not follow the national curriculum. As the access to PIPS is not cost-free, schools have to use their school budgets. All participation is voluntary, although some schools are strongly encouraged to participate by their Local Authorities. In some cases, Local Authorities also get direct access to the data of their schools, if they have paid for the assessments. They are not allowed to make these results public and are supposed to use the data for supporting schools. The feedback is delivered via regular mail (to the PIPS coordinator on the school) and via a secured electronic portal. Depending on whether the assessments were computer delivered or paper based, feedback production can take between 2 days and 8 weeks. The main function of PIPS is a pupil monitoring system (Tymms & Albone, 2002).
- **South African monitoring system for primary schools (SAMP):** PIPS served as a basis for the development of this system. SAMP consists of domain-specific assessments for students aged 5 to 6. It has evolved into a distinct SPFS, developed at the Centre for Evaluation and Assessment at the University of Pretoria (South Africa). Due to resource limitations, feedback with regard to assessment results is only delivered in the Tshwane Region for the 1st year of primary education. Furthermore, only the government-funded schools are reached as these are the schools with the largest need for accessible assessment systems, in contrast to the wealthier independent schools. Therefore, this SPFS delivers feedback for free (limited to 80 learners per school). Very specific to the development of SAMP is the complicated language context of South Africa, with 11 official languages. SAMP is restricted to the three predominant languages of instruction in that region: English, Afrikaans, and Sepedi. Therefore, SAMP is a small-scale SPFS offering feedback to 22 schools. All schools are participating voluntarily. The feedback users are primarily the school team members, and specifically administrative staff. They are free to communicate the results with other stakeholders, such as parents, the department of education, and so forth. Feedback supply via regular mail is not an option as there is no assurance the package will reach its destination in South Africa. Since many schools lack Internet and even computer access, electronic feedback delivery is also not an option. Therefore, feedback is delivered on the school site to the contact person. This happens 4 days to 2 weeks after data gathering (Archer, 2010).
- **Leerling- en onderwijsvolgsysteem (LOVS) [Pupil and educational monitoring system]:** LOVS is in the first place a pupil monitoring system (for pupils aged

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

4–12). Dissimilar to the other systems in this study, LOVS is also an official accountability system (e.g., it is used by the Dutch Inspectorate). LOVS consists of several assessments (e.g., maths, literacy), which students take twice a year during their entire primary school career, allowing school to track student achievement over time. The wide acceptance of LOVS is indicated by a 95% rate of use of at least one of the tests in all elementary schools in The Netherlands, including special needs education. The feedback is provided by a private company. Due to this private character, schools use their own budgets to pay for services offered. As a consequence, they are also the owners of their data. To disseminate results to externals (e.g., other schools), schools need the permission of parents. The way of delivering feedback depends on the tests taken. Some results are sent by regular mail, while other data are provided via an electronic portal, via software on a disk or manually by means of printed scoring tables. Also, depending on the test taken and the standardization process (based on previous or current reference groups), the period for delivery of feedback can take anything from a matter of seconds to a few months.

- **Schoolfeedbackproject (SFP) [School feedback project]:** The SFP is a research and development project, set up by three universities in Flanders (Belgium). As no central assessment system exists, schools lack information about their performance as compared to the national average or to results of schools with similar characteristics. Therefore, a government-funded project has been set up for creating a Flemish SPFS, consisting of domain-specific achievement tests for pupils aged 6 to 12. In this article, only the system developed for primary education will be described and analyzed (Year 1–6). Although SFP participation was on a voluntary base, some central school boards decided on school participation. Assessment results are fed back confidentially to school members only. In addition, aggregated results are reported to the educational authorities, as part of the research project. School reports are delivered to the feedback coordinator on the school by electronic mail. Due to the research and development nature of the SFP project, feedback generation took several months. In the future, feedback will be delivered much faster as the underlying software engine, feedback formats, and reference groups are now available. Furthermore, the SFP is developing a secured electronic portal to download results (Vanhoof, Verhaeghe, Verhaeghe, Valcke, & Van Petegem, 2011).

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

Instruments

A survey was developed to gather information on the different technical characteristics of SPFSs. The questionnaire consisted mainly of multiple-choice items, with some additional open questions. For some items, the respondents were requested to provide complementary explanation for their responses. All different options were summed up and explained in a text file, including 46 items:

- 11 items with background information to identify the SPFS (e.g., geographical delivery area, users, participation of schools);
- 9 items on the data gathering process (e.g., place of data gathering, who gathers the data, data gathering instruments);
- 7 items on the data analysis (e.g., type of analyses, scaling model, statistical model);
- 14 items on the content of the feedback report and the concepts used (e.g., subjects, types of variables, reference groups);

- 5 items on the graphical data representation (e.g., feedback delivery medium, graphical representations used, representation of confidence intervals).

Procedure

The SPFSs survey was sent to the directors or coordinators of the five selected SPFSs. They were informed about the purpose of this study. Additionally, semistructured in-depth telephone interviews were set up, to elaborate or clarify survey answers and to gather information about the rationale underlying certain SPFS characteristics. The telephone conversations, which took on average 90 minutes, were audio-taped with permission of the interviewees and subsequently transcribed. The integrated results from both the survey and the interview were sent to the interviewees for member checking.

Finally, the integrated surveys and interviews data files were summarized for each feedback system. These files were reorganized into a conceptually ordered meta-matrix (Miles & Huberman, 1994) that facilitated a variable-oriented (vertical) and case-oriented (horizontal) analysis. In this matrix, we compared the SPFSs on the following characteristics:

- data gathering (e.g., medium, type of instruments, timing);
- data analysis (e.g., type of analysis, model used, type of value added);
- feedback content (e.g., variables, subjects, reference groups);
- graphical representations (e.g., feedback medium, graphical representations used, reliability indices).

This meta-matrix helped to develop a quick overview of the variety in feedback systems. The meta-matrix helped to visualize the sets of cases, and it helped to clarify case relationships in ways that facilitated comparison. Each case was condensed as such that it permitted a systematic visualization and comparison (Miles & Huberman, 1994). Parts of this meta-matrix will be illustrated and explained in the results section (see Tables 3 to 8).

Results: comparison of the SPFSs

Data gathering

In the context of a SPFS, a clear data gathering process is of major importance to evaluate the accuracy of the data on which the feedback is based. Therefore, the following elements were studied: the persons gathering the data (the data gathering process), types and structure of instruments used, data gathering tools, data sources, and time. The results are summarized in Table 3.

In almost all cases (asTTle, PIPS, LOVS, SFP), teachers and/or other school team members organized the test administration in the school, following strict test protocols. Only in case of SAMP, field workers from the SPFS guided the assessment. The latter choice was made in view of reliability of the data collection and not to interrupt teaching involvement. As for the other SPFSs, teachers not only organized the testing but sometimes also provided data about a pupil's functioning. In PIPS and LOVS, for example, they completed observation scales, pupil background questionnaires, and/or surveys on the socioemotional functioning. In asTTle, teachers have an even more active role by composing the test based on predefined parameters and options by using the testing

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

Table 3. Overview of data gathering process, medium, sources, and timing.

	AsTTle	PIPS	SAMP	LOVS	SFP
Data gathering process	• Teachers compose test and organize test administration • One-on-one testing and collective testing • Test administration mostly in classroom (pupils have own computer); sometimes at home • Flexible test supply • Prescribed rules	• Teachers or school members organize test administration, and complete observations, pupil background questionnaire for children under 9, and ADHD scale • One-on-one testing for younger children; collective testing for older children • Test administration in classroom or in computer lab • Flexible test supply • Prescribed rules	• SAMP field workers organize test administration • One-on-one testing • Test administration at school • Fixed test supply • Demand-driven supply if several schools have same request • Prescribed rules	• Teachers organize test administration, some schools hire externals for test administration • One-on-one testing and collective testing • Test administration in classroom (paper and pencil) and computer lab • Fixed test supply for some tests, flexible test supply for others • Prescribed rules	• Teachers organize test administration, parents complete student background questionnaire • One-on-one testing and collective testing • Fixed test supply • Prescribed rules
Medium	• Parallel paper-pencil, computerized, or computer-adaptive version of test	• Parallel paper-pencil, computerized, or computer-adaptive version of test	• Paper-pencil test	• Parallel paper-pencil and computerized version of test	• Paper-pencil test
Data source	• Pupils age 4/5 to age 16/17	• Pupils age 4/5 to age 10/11	• Pupils age 5/6	• Pupils age 4/5 to 11/12	• Pupils age 6/7 to 11/12
Timing	• At any time: standardizations for different moments in the year	• Fixed moments: standardization related to fixed moment	• Fixed moment: baseline at start of the year and follow-up at the end of year	• Fixed moments: standardization related to fixed moment • At any time: only useful for individual learning paths as standardizations are related to fixed moments	• Fixed moments: standardization related to fixed moment

Table 4. Overview of data gathering instruments used in selected SPFSs.

	asTTle	PIPS	SAMP	LOVS	SFP
Completely structured					
Domain-specific tests	X	X	X	X	X
Survey on attitudes/socioemotional development	X	X	X		
General achievement test		X		X	
Observation scale		X		X	
ADHD scale		X			
Pupil background questionnaire		X			X
Test on study skills				X	
Survey on social emotional functioning				X	
Test of intelligence				X	
Test on interests				X	
Semistructured					
interviews on strategies in mathematics, writing assessments	X				
Pupil background questionnaire			X		
Rating scale for evaluation of a technical piece of work				X	
Computer adaptive					
Domain-specific tests	X	X[1]		X[2]	
Screening instrument for Dyslexia				X	
Other					
Automatic upload of pupil background variables from data management system	X	X		X	
Observation notes of testing: no structured instruments			X		
Upload of results from Statutory Assessment Tests		X			

Notes: [1]Computer-delivered version of PIPS for Years 1–6; all other tests use stopping rules based on a number of mistakes made, on increasingly difficult items. [2]depending on the test taken.

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

software tool. Furthermore, parents can also be asked to provide information. In the case of the SFP, a parent questionnaire was provided for gathering home and pupil background information.

Not only the testing instructions but also most testing instruments are highly structured. Table 4 gives an overview of the instruments used. Almost all instruments are highly prestructured. This means that tests and questionnaires entirely prescribe and guide the data collection. In some cases, semistructured instruments are used. For example, SAMP does not require schools to complete structured questionnaires about student background variables, but only lists what information would be favorable to gather (due to a lack of pupil information and lack of computerized management system). In contrast, asTTle, LOVS, and PIPS make use of advanced software options that allow automatic import of pupil-level data from the school's management information systems. These three SPFSs additionally provide computer adaptive testing. Test items are presented to pupils according to their ability level. For example, if a pupil performs well on an intermediate difficult item, a more difficult question will be presented. Subsequently, if he performs poorly, he will be presented with an easier item.

Testing pupils can be very time consuming, especially in the case of young children. As they do not master reading or writing skills, often one-on-one oral testing is necessary. In this case, the instructor provides the explanation following the protocol and the pupil provides the answers (e.g., in PIPS and SAMP). In other cases, a one-on-one testing is required because of the nature of the test (e.g., reading fluency in SFP and LOVS). In

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

Table 5. Overview of data analysis used in selected SPFSs.

	asTTle	PIPS	SAMP	LOVS	SFP
Scaling model: classical test theory	Interviews and surveys	Surveys, observation scales, and PIPS reception test	All instruments	Surveys, observation scales, and some tests	Test of reading fluency
Scaling model: item response theory	Rasch for tests	Rasch for tests Year 1–6	Rasch for defining item parameters	Technique depends on the test taken	Two-parameter model for all tests, except for reading fluency
Statistical model used	None	Ordinary least squares	None	Multilevel analysis for growth curves	Multilevel analysis with repeated measures for piecewise growth models
Value added	None	– Prior and concurrent – At pupil and aggregated level	None	None	– Contextual – At aggregated level
Levels of unit: respondents (per pupil/class/teacher/subgroup/cohort/year group/school/school network/district/province/state)	– All levels from individual pupils to school – Sometimes school network level	– All levels from pupil to year group	– Pupil – Cohort/school – Group: all schools using the same language of instruction	– All levels from individual pupils to school	– Cohort/year group – School
Levels of unit: content (per item/subscale/subject/aggregation/educational goal)	– Subscale – Subject	– Item – Subscale – Subject – Aggregate: general achievement score	– Subscale – Subject – Aggregate: general achievement score	– Subscale – Subject – Aggregate: general achievement score	– Subscale – Subject

Note: In SFP, contextual value added is based on adjustment both for student background and prior achievement; in PIPS, concurrent value added is based on adjustment for developed ability (picture vocabulary and nonverbal cognitive skills) but not prior achievement, but is also labeled context value added.

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

Table 6. Overview of feedback content.

	AsTTle	PIPS	SAMP	LOVS	SFP
Type of variables	• Attitudinal • Behavioral • Cognitive • Contextual	• Attitudinal • Behavioral • Cognitive • Contextual	• Attitudinal • Behavioral • Cognitive • Contextual	• Attitudinal • Behavioral • Cognitive • Contextual	• Cognitive • Contextual
Subjects	• English language • Mathematics • Writing • Maori: Panui, Tuhituhi, Pangarau	• Language • Mathematics • Science	• Language • Mathematics • Foreign language: English	• Language • Mathematics • Foreign language: English • World orientation (geography, history, and environmental science)	• Language • Mathematics
Other information	• School culture • Pupil background • Pupil mobility	• Pupil background • Mathematics, language and science attitude • School attitude • Socioemotional development • ADHD scale • Picture vocabulary and nonverbal ability	• Pupil background • Mathematics and reading attitude • School attitude • Handwriting	• Pupil background • Socioemotional development • Study skills • Dyslexia • Intelligence • Space and time • Sorting • Development of preschool children	• Pupil background • Pupil mobility
Test goals	• National curriculum	• National curriculum	• Skills based, but linked to curriculum	• National curriculum	• National curriculum

(*continued*)

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

Table 6. (*Continued*).

	AsTTle	PIPS	SAMP	LOVS	SFP
Reference group	• National average • School similar as mine: based on 5 indicators • Self-referenced • Norm referenced • Criterion referenced	• Representative sample of population • Self-referenced • Norm referenced	• Group of participating schools using the same medium of instruction; no comparison across groups • Self-referenced • Norm referenced • Criterion referenced	• Representative sample of population • Comparison with specific subgroups based on pupil background variables • Self-referenced • Norm referenced	• Representative sample of population • Top-5 and bottom-5 scoring schools • Self-referenced • Norm referenced
Text	• Adapted to school • Results and how to interpret	• Standardized text • Results and how to interpret	• Adapted to school • Only results	• Standardized in general • Adapted to school in some cases • Only results	• Standardized • Results and how to interpret

Table 7. Numerical measures used in selected SPFSs.

	asTTle	PIPS	SAMP	LOVS[1]	SFP
Type of scores					
Adjusted scores	X	X	X	X	X
Raw scores	X	X	X	X	X
Performance indicators					
Band score		X		X	
Cut-off score	X		X		
Grade score		X		X	
Learning gain score	X	X	X[2]	X	X
Mean score	X		X	X	X
Percentage score			X	X	
Percentile score				X	
Rescaled score			X	X	X
Standardized score		X		X	
Value-added score		X			X

Note: [1]Depending on the tests taken; [2]SAMP also registers loss scores besides gain scores.

Table 8. Overview of representation modes in selected SPFSs.

	asTTle	PIPS	SAMP	LOVS[1]	SFP
Medium: fixed					
Printed report		X	X	X	
PDF version		X	X		X
Medium: flexible tools					
Online tools	X				
Software applications on local network	X			X	
Excel sheet			X		
Excel macro's in sheet		X			
Graphical representations					
Bar graph	X		X	X	
Box plot	X	X			
Cross table	X	X	X	X	X
Divided bar graph	X	X		X	
Grouped bar graph				X	
Histogram			X		
Layer graph				X	
Line graph	X			X	X
Multipanel display				X	
Pie graph				X	X
Scatter plot with regression line		X			
Side by side graph				X	
Other: e.g., schemes, iconic representations	X				X
Reliability indices					
Confidence intervals	X	X		X	
Significance values					X

Note: [1]Depending on the tests taken.

other cases, one-on-one testing is optional, as the testing medium allows both individual and group-based testing. This is the case for asTTle, as each pupil owns a personal computer and software adapts standardization of the scores to the moment of testing.

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

More rigid systems with paper-pencil tests, computerized tests in computer labs, and/or fixed measurement moments are more likely linked to whole classroom testing (PIPS, LOVS, SFP).

The place of testing is highly related to the school infrastructure. Mostly, tests take place in the classroom if printed booklets are used (asTTle, PIPS, LOVS, SFP), or in the computer lab for computerized versions (PIPS, LOVS). Testing administration of asTTle is flexible because of technological provisions. Even testing at home is possible. In the case of SAMP, each testing situation is slightly different as – at the local level – an appropriate place is being looked for (e.g., in the staff room, under a shady tree).

Data analysis

In this section, we focus on the underlying scaling model used, on the data analysis model used including value-added measures, on the opportunities for longitudinal measurements, on the inclusion of pupil mobility, and on the aggregation levels being adopted (see Table 5). Being informed about the data analysis of SPFSs is a prerequisite for making judgments about feedback accuracy. In all feedback systems, data are analyzed quantitatively. We will focus in this section on the diversity in analysis techniques.

First, the underlying scaling models have been examined. Item response theory (IRT) is underlying all SPFSs; but with varying degrees. IRT is based on the notion that the probability of a correct response to an item can be modeled as a function of both the respondent's ability level and the item difficulty. This technique thus estimates several parameters, including the difficulty level of the items and the ability scores of the respondents (for more information, see Baker, 2001; Van der Linden & Hambleton, 1996). By creating one skill scale that relates different tests in a certain domain, IRT offers opportunities for longitudinal measurements or computer-adaptive testing. IRT has been applied in the selected SPFSs for defining the item parameters and composing tests (SAMP). AsTTle, PIPS, LOVS, and SFP go further and use IRT for defining ability test scores for the respondents for certain test versions. The IRT model that has been used most widely is Rasch (in asTTle, PIPS, and SAMP). The techniques used in LOVS depend on the test taken, and SFP uses a more complex two-parameter model. The system taking the most advantage of IRT is asTTle. In combination with software tools, teachers are supported to compose tests from an item bank with different degrees of difficulty. Besides IRT, classical test theory (CTT) is applied in all systems. This is not only used for analyzing data from interviews, surveys, and/or observation scales (asTTle, PIPS, LOVS) but also for some tests (SAMP, SFP) which require no further analysis than the calculation of sum scores.

Only PIPS and SFP make explicit use of value-added measures. These measures give an indication of what the school has added to the learning process of its learners (Mortimore, Sammons, & Thomas, 1994; Van de Grift, 2009). Generally speaking, scores on value-added measures in education reflect the difference between an observed outcome (most often a test score) and the outcome that would be "normal" for a student with the same background characteristics (such as gender, ethnicity, family income, prior achievement, and cognitive aptitudes). The "normal" scores may be obtained through fairly advanced statistical methods, such as regression analysis or multilevel modeling. In other cases, straightforward group means (e.g., the average math score by gender) may serve as a reference basis. Scores on a value-added measure thus express to what extent an individual scores higher or lower than similar counterparts. The principle can be applied both to individual students and to groups (e.g., classes, schools). The observed scores are

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

often called raw scores, and the "normal" scores are usually referred to as adjusted scores. The basic idea behind the use of value-added measures is to compare like with like.

PIPS makes a distinction between prior and concurrent value added. Prior value-added measures relate to student achievement scores measured against prior achievement, while concurrent (or context) value added relates to achievement measured against a developed ability score. Developed ability is assessed by means of a combination of language acquisition (picture vocabulary) and a nonverbal test. This information will be useful in situations where prior achievement scores were particularly high or low because of very effective or ineffective teaching in the past. In contrast to PIPS, SFP uses both student background variables and prior achievement scores in the estimation of contextual value added. Both systems conflict in this conception as student background variables are either seen as not necessary to include or as necessary variables to be included in the model. Furthermore, the value-added approaches differ significantly in their level of reporting. While SFP is convinced that value added should only be reported at an aggregate level, PIPS allows pupil-level residual analysis. LOVS implicitly applies the notion of value-added measures by reporting the difference in growth of the school as compared to the reference group.

When focusing on the statistical model, underlying the feedback production, a large variety in complexity can be noticed. While some SPFSs strive for complexity to provide a nuanced view on school performance data (as SFP and LOVS), others consciously avoid complexity in favor of transparency for feedback users. For example, in the calculation of value added, PIPS applies an ordinary least squares in contrast to the multilevel piecewise growth curve models of SFP. Other systems do not use regression models as they do not intend to calculate value-added scores in order to keep the low-stakes character of testing (asTTle) or are still in a development phase (SAMP).

The type of statistical models used affects the options for longitudinal measurement. This means that scores for pupils are linked to each other over time. Whether or not learning progress can be measured depends on the scale used. In case of asTTle, progress is estimated on one underlying IRT ability scale, linking all tests in a certain domain. PIPS and LOVS use a scale of standardized scores (either obtained by CTT or IRT) and put these scores on a time line. SFP in contrast not just places the (rescaled) IRT scores on a time line but adjusts these scores both for the influence of prior achievement and for pupil background characteristics by building on a repeated measures model. These adjusted scores do not express the actual achievement level, but the level that would have been achieved if the pupils have comparable background characteristics as a reference group. This results in a different conception of growth and longitudinal measurement.

Another factor delineating opportunities for longitudinal analysis is the number of measurement occasions. AsTTle, PIPS, LOVS, and SFP offer tests with (at least) three linked measurement moments, while SAMP only tests pupils at the start and the end of the 1st year of primary education. In all systems, the users decide whether or not to participate in single or successive measurements.

Moreover, it is of importance to stress the influence of pupil mobility, in particular when longitudinal data are represented for a cohort. The additional value of asTTle – as it does not adopt a value-added approach – is that they report student-level data for all students, irrespective what schools they have attended before. There is no need to match data as there would be in a value-added approach. In more complex longitudinal models, taking into account pupil mobility requires cross-classifications, which can (over)burden the statistical analysis capabilities. Another consequence of pupil mobility is that values

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

are missing for pupils that left a testing sample. A solution is to apply repeated measures to simulate data from missing values, as in the SFP.

A final aspect to be discussed in this section is the reported aggregation level for respondents and content. With regard to the respondents, all systems opt for reporting pupil-level data, with the exception of the SFP, which is designed for evaluating and informing school policy with a focus on aggregated data. The systems that report pupil-level data (asTTle, PIPS, SAMP, and LOVS) also report data on aggregated levels as classroom level, group level, school level, and so forth. All SPFSs report at (broad) subscale and subject level. Only PIPS reports at item level (only for reception feedback) as this can inform classroom planning. AsTTle intentionally does not report at item level since this could lead to "teaching to the test". Another restriction for reporting item-level scores depends on the objective of the test taken. SFP uses tests for determining learning achievement (thus requiring to avoid ceiling effects) and not for diagnostics (necessity to determine outliers). Therefore, it is less appropriate to report item-level scores.

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

Feedback content

This paragraph contains a description of the subjects and topics that are reported in the feedback reports, the conceptual representations (performance indicators) and reference groups used, and the sections offered in reporting. Following the quality standards for performance indicators (Visscher & Coe, 2002, 2003), the feedback content has to be relevant and useful. Furthermore, SPFS users should be willing/able to accept the performance indicators and consider them to be fair. An overview of the content of the feedback can be found in Tables 6, 7, and 8.

Regarding the content being tested in the selected SPFSs, we refer to Table 4, in which the data gathering instruments are summarized, and Table 6. These tables show that all systems build on domain-specific tests; in all cases language and mathematics tests consisting of different subscales. Some systems broadened the test range with science tests (PIPS), English as a foreign language (SAMP, LOVS), and/or technology and world orientation (aggregation of geography, history, and environmental science in LOVS).

The other data collection instruments reported in Table 4 focus on noncognitive measures, such as attitudinal, behavioral, and contextual contents. As to behavioral scales, PIPS, for example, offers a scale for detecting ADHD and LOVS for dyslexia, whilst handwriting is tested by asTTle and SAMP. To tackle attitudinal measurements, there are measures of attitudes related to subjects (asTTle, PIPS, SAMP, and LOVS), to the school culture in general (asTTle, PIPS, and SAMP), or to socioemotional development (LOVS). Related to contextual information, informing schools about pupil mobility seems to be of importance to develop an understanding of their functioning. Data with regard to why pupils are leaving, which newcomers schools are attracting, which pupils go to special education, and the number of pupils with learning delays can stimulate reflection at the school level, which can transcend individual learning trajectories. Only the SFP reports specifically about the latter.

Numerical measures

A wide range of numerical measures have been reported in the SPFSs in this study. Table 7 gives an overview.

In addition to raw scores, types of adjusted scores (e.g., for prior achievement) are fed back in all cases, although the degree of statistical sophistication may differ. In the

simplest case, the adjusted score is the average for a reference group. In other cases, the adjustments are based on regression analysis or repeated-measures models.

All these types of scores are rescaled into meaningful units for the users. For example, scales are created with a mean of 50 and a standard deviation of 15. All these transformations are somehow arbitrary as there are no conventions as to which scales, bands, or grades are to be favored. Mostly, test scores have been transformed in view of the local context. For example, AsTTle and PIPS reformulate scores to grades in accordance to the national curriculum, SAMP rescales to 5-point scales teachers are familiar with, and LOVS expresses scores in line with preferences of the inspection authorities.

Feedback reports may contain more information than the mere test results. The explanation on how to interpret the results is only provided in the feedback reports of PIPS and SFP. Other systems provide this information in an accompanying manual. When it comes to searching for explanations for the results for a specific school, no further help is provided in any report. However, AsTTle, LOVS, and SAMP take considerable initiatives for offering remediation material. AsTTle is the most advanced system by offering supporting material for teachers in accordance to the achieved grade levels per pupil and group.

What information can be derived from the reports also depends on the references offered (norm, self-, or criterion reference). These three reference forms offer different opportunities for a school to compare their own functioning. All systems offer a *norm* to compare results. In most cases, this reference builds on (a representative sample of) the national average. SAMP cannot work in this way due to the small scale of the local project. Instead, SAMP offers the opportunity to compare with same language schools within the sample. AsTTle and LOVS allow comparing with comparable schools, based on particular characteristics. These features foster fair comparison, but build on different calculation procedures for adjusted scores. Opportunities for *self-reference* are offered in all systems by allowing schools to compare results over time, either within cohorts (cf. gain scores, longitudinal measurements) or between cohorts (multiple measurements with different year groups). *Criterion-based references* are less prevalent (asTTle and SAMP) as these imply an absolute instead of relative point of reference. In these cases, cut-off scores are often used. SPFS do not provide these cut-off scores. It is up to users to determine these cut-off scores. These cut-off scores reflect user perspectives on expected minimal performance levels or values.

Graphical representations

With regard to representation modes used in the feedback reports, we discuss the medium used to present the results, the graphical representations, and the attention paid to reliability indices (see Table 8).

SPFSs differ in the feedback media used to report the results. These media are related to the flexibility for users in choosing representations or ways to manipulate the feedback output. In AsTTle, PIPS, and LOVS, users can select different types of representations through software tools or Excel macros. They can, for example, select a table to present exact data, and growth curves to show trends. SAMP and SFP are less flexible: These SPFSs provide the user with a printed or digital PDF report of the results. SAMP additionally reports the results in Excel sheets, which users can use to carry out secondary analyses.

The graphical representations offered by the different SPFSs diverge to a large extent. Some systems only include simple representations, such as bar graphs, cross tables, and

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

histograms (SAMP). Others include complex graphical representations of the results, such as scatter plots with superimposed regression lines (PIPS), line graphs (SFP), and layer graphs (LOVS).

The school performance feedback results are based on varying statistical analyses. To enable users to judge the accuracy and importance of their findings, information about the "uncertainty" of the results has been incorporated in asTTle, PIPS, LOVS, and SFP. This is often done by adding confidence intervals to the results. All SPFSs studied present confidence intervals via either bar graphs (AsTTle and LOVS) or longitudinal progress charts (PIPS). SFP represents uncertainty by marking significant values in cross tables. SAMP prefers not to present confidence intervals, as this would result in a too complex interpretation/representation of the results. Instead, they warn the users not to overinterpret small differences or small shifts in scores.

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

Conclusion and discussion

Goal of the study

As data-based decision making is receiving increased attention in education, more and more SPFSs are being developed and used worldwide. However, little research is available about the characteristics of the different SPFSs. It is important to consider the SPFS characteristics when developing or selecting one to use. Users need to purposefully choose an SPFS corresponding to their information needs. This requires a transparent view on SPFS characteristics. Therefore, in this article, we compared the technical characteristics of SPFSs. We illustrated diversity in the data gathering processes, the type of analyses, and the content of the feedback, including the numerical measures and representation modes used. The goal of this study hereby was not to judge the quality of the different SPFSs but to highlight SPFS characteristics.

Data gathering

With regard to data gathering, all SPFSs studied mainly present completely prestructured instruments, as cognitive tests, questionnaires on socioemotional development, building on diverse scales types. Semistructured instruments such as interviews or rating scales are provided as well. All instruments are accompanied by protocols on how to gather the data. Providing such highly structured instructions and instruments is a prerequisite for a standardized and reliable data collection (Fitz-Gibbon, 1996; Fitz-Gibbon & Tymms, 2002), especially if data will be gathered by school staff. As data collection is very time consuming, technology-supported tools present advantages. Therefore, initiatives such as computer adaptive testing or automatic upload of data from management information systems (as in asTTle, PIPS, and LOVS) facilitate efficient data collection. These tools take away part of the burden of pupils and teachers during data collection, and foster targeted data collection. However, only advanced SPFSs currently present such tools. Also, these software tools cannot be used in contexts with a weaker infrastructure.

Feedback content

With respect to the content of the feedback, the SPFSs in this study adopt a narrow focus on cognitive outcomes (e.g., language, mathematics, and/or science) that are part of the core curriculum. Developers of SPFSs might consider including other school subjects, as

well as instruments to collect attitudinal, behavioral, and contextual information. If schools want to make informed decisions on how to improve their education, they need to build on different types of data (Schildkamp & Kuiper, 2010; Schildkamp & Lai, 2013). AsTTle, PIPS, and LOVS seem to move into this direction, but also other types of data could be considered, such as data about the functioning of teachers (e.g., teacher and student questionnaires). A preferable scenario to foster the resulting data triangulation is the development of integrated management information systems (Bosker, Branderhorst, & Visscher, 2007). In order to obtain an integrated system, more coherence in data conceptualization and representation is required, not only between different data sources but also between different instruments of the same SPFS. A first step is that SPFS developers adopt a larger conformity in data analyses and data representations.

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

Data analysis

With regard to data analysis, it is important to find a balance between statistically correct – and often complicated – analyses and accurate results, on the one hand, versus understandable analyses and user-friendly results, on the other hand. For example, the analyses used in PIPS are fairly straightforward and not too complex. Schools can understand the results, and studies show that schools subsequently feel ownership of the results (Tymms & Albone, 2002), which directly influences the degree to which the feedback is actually used (Kyriakides & Campbell, 2004; Schildkamp & Teddlie, 2008). However, because no multilevel analyses have been applied, schools are sometimes wrongly classified as underperforming. To reduce misclassifications, researchers claim that it is critical always to apply multilevel models (Goldstein & Spiegelhalter, 1996; Karsten, Visscher, Dijkstra, & Veenstra, 2010). Building on the observations of Yang et al. (1999), it seems to be possible to explain even these complex multilevel models and outcomes to head teachers. In contrast, others consider multilevel modeling as inappropriate for feedback purposes and claim that the method of ordinary least squares is sufficiently accurate and understandable (Fitz-Gibbon, 1996; Fitz-Gibbon & Tymms, 2002; Sharp, 2006). Whatever statistical analyses are being adopted, it should inform its users about the related constraints. As stated by Hellrung and Hartig (2013), it is crucial that users are able to understand the data from the SPFS before they can actually use it and before it can result in better student learning. So, instead of choosing transparency and interpretability over using more sophisticated analyses methods, it might be considered to invest in developing the staff capacities in understanding data and using the data from these SPFSs.

Moreover, it is important to communicate that every measurement reflects some type of error. Statistical estimates always include uncertainty, which needs to be taken into account during interpretation. This is especially true when building on data from small groups, classes, and schools. A SPFS should therefore provide information about its limitations and uncertainties, and provide information about the reliability of the estimates (Fitz-Gibbon & Tymms, 2002; Goldstein & Myers, 1996; Goldstein & Spiegelhalter, 1996; Karsten et al., 2010; Mortimore et al., 1994; Rowe, 2004; Yang et al., 1999). The importance of the quality of the data also depends on the purpose(s) of data use. For example, if the stakes are high, information about the reliability is essential; the data need to be of high quality (Coburn & Turner, 2011; Schildkamp & Kuiper, 2010). If the stakes are low, for example, the data are used for formative assessment for learning activities, the quality of the data is perhaps slightly less important. The latter data are in this case used to (re)direct learning processes. If the changes in the learning environment made based on

these data do not produce the intended effects, this will become quickly clear from the next assessments, whereupon new changes can be made.

If school-level data are used for making comparisons with reference groups, the systems can make use of value-added measures. Value added is usually defined as everything the pupil has learned at his/her school (e.g., Van de Grift, 2009). However, the concept "value added" is not unproblematic (Van de Grift, 2009). It is not possible to assess everything a pupil has learned, such as social and creative abilities. Furthermore, because pupils change schools and classes, different schools and classes influence pupils' progress. Also, it is not clear how learning progress should be measured, and how to take into account the knowledge and skills acquired outside the schools. As a result, several problems have been associated when applying value-added modeling (Karsten et al., 2010; Van de Grift, 2009), such as:

- the problem of missing values, which may distort the results. Missing data might not be random, but might result from interventions in schools. Incorporating the impact of missing values in the estimation procedures is therefore advisable (Sanders, 2006; Van de Grift, 2009; Yang et al., 1999).
- the instability of value-added judgments. It is therefore recommended to build on data from successive cohorts (at least 3 school years; Van de Grift, 2009) and to use longitudinal measurements (Heck, 2006) or to build on average scores from successive years (Organisation for Economic Co-operation and Development [OECD], 2008).
- there are different procedures for computing value-added models, resulting in a different ranking of schools (Fitz-Gibbon, 1996; Goldstein & Spiegelhalter, 1996; Heck, 2006; OECD, 2008; Rowe, 2004; Sanders, 2006; Van de Grift, 2009; Yang et al., 1999). For example, no consensus exists about the inclusion/exclusion of student background characteristics in the SPFS models. As student achievement results are influenced by prior achievement and student background characteristics (such as gender and socioeconomic status [SES]), several researchers suggest correcting for these extra scholar influences (Goldstein & Myers, 1996; Goldstein & Spiegelhalter, 1996; Goldstein & Thomas, 1996; Heck, 2006; Karsten et al., 2010; Rowe, 2004; Sanders, 2006; Yang et al., 1999).
- the predictive validity of a value-added model remains limited for certain schools (e.g., for schools with large SES-gaps). SPFSs using value-added models should therefore always be careful when categorizing schools as underperforming, and should adopt labels such as durably underperforming or durably outperforming instead of ranking schools (Van de Grift, 2009). Goldstein and Thomas (1996) and Yang et al. (1999) also recommend using this procedure only to identify "institutions at extremes", as a screening device to detect problems.
- users have difficulties when interpreting value-added data (Karsten et al., 2010; Santelices & Taut, 2009; Vanhoof et al., 2011). Users should be supported to acquire expertise in data interpretation by, for example, getting offered more and diverse value-added models (Schatz, VonSecker, & Alban, 2005).

Graphical representations

After having analyzed the data, SPFS developers need to carefully consider what types of numerical measures and graphical representations are offered to users. Research revealed that even simple numerical conceptions and representations are often interpreted

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

incorrectly. Teachers' statistical knowledge is often insufficient (Earl & Fullan, 2003; Hellrung & Hartig, 2013; Zupanc et al., 2009). A sufficient level of assessment literacy is a prerequisite for a correct understanding. Mandinach (2012) goes further and states that teachers need pedagogical data literacy: the ability to analyze data and, based on the data, combined with pedagogical content knowledge, take meaningful action. If not, proper support initiatives should be foreseen.

SPFS developers should keep in mind that the use of school performance feedback does not always lead to improvement, and at least should not be harmful (Fitz-Gibbon & Tymms, 2002; Rowe, 2004). Moreover, they should consider offering training in the interpretation and use of the results, especially when adopting more advanced statistical modeling. Research clearly reveals that SPFS usage without proper training is difficult (Schildkamp & Visscher, 2009; Verhaeghe et al., 2010; Vanhoof et al., 2011). This professional development should include training with regard to data collection and analysis, and, perhaps even more important, how to connect data to the daily practice of school leaders and teachers (Black & Wiliam, 1998; Datnow et al., 2007; Supovitz & Klein, 2003).

Using data from SPFSs

Finally, we want to stress that high-quality SPFSs that give schools opportunities to collect and analyze the feedback data is but a first step towards effective data-based decision making. It is merely a pre-condition (Downey & Kelly, 2013; Schildkamp & Lai, 2013; Wayman, Spikes, & Volonnino, 2013). Subsequent steps of SPFS usage that go beyond the data analysis and interpretation are also difficult; for example, how to identify appropriate measures based on data (Marsh, Sloan McCombs, & Martorell, 2010). Studies show (e.g., Marsh, 2012) that schools need support in all the steps related to data-based decision making: gathering data, analyzing data, combining information with expertise and understanding to build knowledge, knowing how to respond and take action based on data, and assessing the effectiveness of the outcomes that result from the actions taken. Only when teachers and school leaders are involved in all these steps, feedback data from SPFSs will actually lead to increased student learning.

Acknowledgements

We would like to express our sincere gratitude to the directors and researchers of the SPFSs involved in this study for their cooperation: Dr. Christine Merrell, Director of Primary Systems, Centre for Evaluation and Monitoring (CEM), Durham University; Elizabeth Archer, Project coordinator of SAMP, Centre for Evaluation and Assessment, University of Pretoria; Geert Evers, Information Manager Primary Education, Centraal Instituut voor Toetsontwikkeling; Ilse Papenburg: Training and advice, Centraal Instituut voor Toetsontwikkeling; Dr. Jean Pierre Verhaeghe, Project coordinator of the SFP, Ghent University and Katholieke Universiteit Leuven; and Prof. John Hattie, Director of Visible Learning Labs, director of asTTle, University of Auckland

Funding

Part of this research was supported by the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT) [grant number SBO 50194]. IWT is a governmental institute in Flanders, stimulating and supporting society-relevant research.

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

Notes on contributors

Goedele Verhaeghe's PhD was on the use of school performance feedback, and more specifically, the influence of content and representation modes on feedback interpretation. Currently, she is working at the Ministry of Education in Brussels.

Dr. Kim Schildkamp is an associate professor in the Faculty of Behavioural, Management and Social Sciences of the University of Twente. Kim's research, in The Netherlands but also in other countries, focuses on "data-based decision making for school improvement". She has been invited as a guest lecturer and keynote speaker at several conferences and universities, including AERA, the University of Pretoria in South Africa, and the University of Auckland in New Zealand. She is a board member of the International Congress for School Effectiveness and Improvement (ICSEI) and chair of the ICSEI data use network. She has published widely on the use of data.

Prof. Dr. Hans Luyten is an associate professor of education at the Faculty of Behavioural, Management and Social Sciences of the University of Twente at Enschede, The Netherlands. He is an internationally recognized expert on multilevel analysis. His research interests include longitudinal studies both at the individual student level (growth curve analysis) and higher (trends at school and system level), international comparisons, educational disadvantage, and the development of methodologies for assessing the effect of schooling on student development.

Prof. Dr. Martin Valcke is head of the Department of Educational Studies at the Ghent University (Belgium). His research field is the innovation of higher education and performance indicator studies. Detailed info about his publications can be found via http://users.ugent.be/~mvalcke/CV/CVMVA.htm

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

References

Archer, E. (2010). *Bridging the gap: Optimising a feedback system for monitoring learning performance*. Pretoria: University of Pretoria.

Baker, F. (2001). *The basics of item response theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland.

Black, P., & Wiliam, D. (1998). Assessment and classroom living. *Assessment in Education: Principles, Policy, and Practice*, *5*, 7–74.

Bosker, R. J., Branderhorst, E. M., & Visscher, A. J. (2007). Improving the utilisation of management information systems in secondary schools. *School Effectiveness and School Improvement*, *18*, 451–467.

Breiter, A., & Light, D. (2006). Data for school improvement: Factors for designing effective information systems to support decision-making in schools. *Educational Technology & Society*, *9*(3), 206–217.

Campbell, C., & Levin, B. (2009). Using data to support educational improvement. *Educational Assessment, Evaluation and Accountability*, *21*, 47–65.

Carlson, D., Borman, G. D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis*, *33*, 378–398.

Chen, E., Heritage, M., & Lee, J. (2005). Identifying and monitoring students' learning needs with technology. *Journal of Education for Students Placed at Risk*, *10*, 309–332.

Coburn, C. E., Toure, J., & Yamashita, M. (2009). Evidence, interpretation, and persuasion: Instructional decision making in the district central office. *Teachers College Record*, *111*, 1115–1161.

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research and Perspectives*, *9*, 173–206.

Coburn, C. E., & Turner, E. O. (2012). The practice of data use: An introduction. *American Journal of Education*, *118*, 99–111.

Coe, R., & Visscher, A. J. (2002). Drawing up the balance sheet for school performance feedback systems. In R. Coe & A. J. Visscher (Eds.), *School improvement through performance feedback* (pp. 221–254). Lisse: Swets & Zeitlinger Publishers.

Datnow, A., Park, V., & Wohlstetter, P. (2007). *Achieving with data: How high-performing school systems use data to improve instruction for elementary students*. San Francisco: Center on Educational Governance University of California.

Davies, D., & Rudd, P. (2001). *Evaluating school self-evaluation* (Research Report No. 21). Berkshire: National Foundation for Educational Research, Local Government Association.
Downey, C., & Kelly, A. (2013). Professional attitudes to the use of data in England. In K. Schildkamp, M. K. Lai., & L. Earl (Eds.), *Data-based decision making in education: Challenges and opportunities* (pp. 69–89). Dordrecht: Springer.
Earl, L., & Fullan, M. (2003). Using data in leadership for learning. *Cambridge Journal of Education*, *33*, 383–394.
Fitz-Gibbon, C. T. (1996). *Monitoring education: Indicators, quality and effectiveness* London: Cassell.
Fitz-Gibbon, C. T., & Tymms, P. (2002). Technical and ethical issues in indicator systems: Doing things right and doing wrong things. *Education Policy Analysis Archives*, *10*(6). Retrieved from http://epaa.asu.edu/ojs/article/view/285/411
Goldstein, H., & Myers, K. (1996). Freedom of information: Towards a code of ethics for performance indicators. *Research Intelligence*, *57*, 12–16.
Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *159*, 385–443.
Goldstein, H., & Thomas, S. (1996). Using examination results as indicators of school and college performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *159*, 149–163.
Hattie, J. A. C., Brown, G. T. L., & Keegan, P. J. (2003). A national teacher-managed, curriculum-based assessment system: Assessment tools for teaching and learning (asTTle). *International Journal of Learning*, *10*, 771–778.
Hattie, J. A. C., Brown, G. T. L., Keegan, P. J., MacKay, A. J., Irving, S. E., Patel, P., & Campbell, A. R. T. (2005). *Assessment tools for teaching and learning (asTTle) Version 4, 2005: Manual*. Wellington: University of Auckland/Ministry of Education/Learning Media.
Heck, R. (2006). Assessing school achievement progress: Comparing alternative approaches. *Educational Administration Quarterly*, *42*, 667–699.
Hellrung, K., & Hartig, J. (2013). Understanding and using feedback – A review of empirical studies concerning feedback from external evaluations to teachers. *Educational Research Review*, *9*, 174–190.
Hofman, R. H., Dijkstra, N. J., & Hofman, W. H. A. (2009). School self-evaluation and student achievement. *School Effectiveness and School Improvement*, *20*, 47–68.
Jansen, J. D. (2001). On the politics of performance in South African education: Autonomy, accountability and assessment. *Prospects*, *31*, 553–564.
Karsten, S., Visscher, A. J., Dijkstra, A. B., & Veenstra, R. (2010). Towards standards for the publication of performance indicators in the public sector: The case of schools. *Public Administration*, *88*, 90–112.
Kerr, K. A., Marsh, J. A., Ikemoto, G. S., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvements: Actions, outcomes, and lessons from three urban districts. *American Journal of Education*, *112*, 496–520.
Kyriakides, L., & Campbell, R. J. (2004). School self-evaluation and school improvement: A critique of values and procedures. *Studies in Educational Evaluation*, *30*, 23–36.
Lai, M. K., McNaughton, S., Timperley, H., & Hsiao, S. (2009). Sustaining continued acceleration in reading comprehension achievement following an intervention. *Educational Assessment, Evaluation and Accountability*, *21*, 81–100.
Lai, M. K., & Schildkamp, K. (2013). Data-based decision making: An overview. In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data-based decision making in education: Challenges and opportunities* (pp. 9–21). Dordrecht: Springer.
Maier, U. (2010). Accountability policies and teachers' acceptance and usage of school performance feedback - A comparative study. *School Effectiveness and School Improvement*, *21*, 145–165.
Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist*, *47*, 71–85.
Mandinach, E. B., & Honey, M. (Eds.). (2008). *Data-driven school improvement. Linking data and learning*. New York, NY: Teachers College Press.
Mandinach, E. B., & Jackson, S. S. (2012). *Transforming teaching and learning through data-driven decision making*. Thousand Oaks, CA: Corwin.

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

Marsh, J. A. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record 114*, 1–48.
Marsh, J. A., Sloan McCombs, J., & Martorell, F. (2010). How instructional coaches support data-driven decision making. Policy implementation and effects in Florida middle schools. *Educational Policy*, *24*, 872–907.
Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks, CA: Sage.
Mortimore, P., Sammons, P., & Thomas, S. (1994). School effectiveness and value added measures. *Assessment in Education: Principles, Policy & Practice*, *1*, 315–332.
Organisation for Economic Co-operation and Development. (2008). *Measuring improvements in learning outcomes: Best-practices to assess the value-added of schools* Paris: Author.
Park, V., & Datnow, A. (2009). Co-constructing distributed leadership: District and school connections in data-driven decision making. *School Leadership & Management*, *29*, 477–494.
Rowe, K. (2004, April). *Analysing and reporting performance indicator data:* "*Caress" the data and user beware!* Paper presented at the Public Sector Performance and Reporting Conference, Sydney.
Rowe, K., & Lievesley, D. (2002, April). *Constructing and using educational performance indicators*. Paper presented at the Asia-Pacific Educational Research Association, Melbourne.
Sanders, W. L. (2006, October). *Comparisons among various educational assessment value-added models*. Paper presented at The Power of Two – National Value-Added Conference, Columbus, OH.
Santelices, V., & Taut, S. (2009, September). *Comprehension and use of value-added school performance indicators reported to teachers and parents*. Paper presented at the European Conference on Educational Research, Vienna.
Schatz, C. J., VonSecker, C. E., & Alban, T. R. (2005). Balancing accountability and improvement: Introducing value-added models to a large school system. In R. Lissitz (Ed.), *Value added models in education: Theory and applications* (pp. 1–18). Maple Grove, MN: JAM Press.
Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education*, *26*, 482–496.
Schildkamp, K., & Lai, M. K. (2013). Conclusions and a data use framework. In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data-based decision making in education: Challenges and opportunities* (pp. 177–192). Dordrecht: Springer.
Schildkamp, K., & Teddlie, C. (2008). School performance feedback systems in the USA and in the Netherlands: A comparison. *Educational Research and Evaluation*, *14*, 255–282.
Schildkamp, K., & Visscher, A. J. (2009). Factors influencing the utilisation of a school self-evaluation instrument. *Studies in Educational Evaluation*, *35*, 150–159.
Sharkey, N. S., & Murnane, R. J. (2006). Tough choices in designing a formative assessment system. *American Journal of Education*, *112*, 572–588.
Sharp, S. (2006). Assessing value-added in the first year of schooling: Some results and methodological considerations. *School Effectiveness and School Improvement*, *17*, 329–346.
Supovitz, J. A., & Klein, A. (2003). *Mapping a course for improved student learning: How innovative schools systematically use performance data to guide improvement*. Philadelphia: Consortium for Policy Research in Education, University of Pennsylvania Graduate School of Education.
Taylor, J. (2007). The usefulness of key performance indicators to public accountability authorities in east Asia. *Public Administration and Development*, *27*, 341–352.
Tymms, P. (1999). *Baseline assessment and monitoring in primary schools*. London: Fulton.
Tymms, P., & Albone, S. (2002). Performance indicators in primary schools. In A. J. Visscher & R. Coe (Eds.), *School improvement through performance feedback* (pp 191–218). Lisse: Swets & Zeitlinger.
Van de Grift, W. (2009). Reliability and validity in measuring the value added of schools. *School Effectiveness and School Improvement*, *20*, 269–285.
Van der Linden, W., & Hambleton, R. K. (Eds.). (1996). *Handbook of modern item response theory.* Heidelberg: Springer.
Vanhoof, J., & Van Petegem, P. (2007). Matching internal and external evaluation in an era of accountability and school development: Lessons from a Flemish perspective. *Studies in Educational Evaluation*, *33*, 101–119.

Downloaded by [Universiteit Twente] at 02:33 11 November 2015

Vanhoof, J., Verhaeghe, G., Verhaeghe, J. P., Valcke, M., & Van Petegem, P. (2011). The influence of competences and support on school performance feedback use. *Educational Studies*, *37*, 141–154.

Verhaeghe, G., Vanhoof, J., Valcke, M., & Van Petegem, P. (2010). Using school performance feedback: Perceptions of primary school principals. *School Effectiveness and School Improvement*, *21*, 167–188.

Visscher, A. J. (2002). A framework for studying school performance feedback systems. In A. J. Visscher & R. Coe (Eds.), *School improvement through performance feedback* (pp. 41–71). Lisse: Swets & Zeitlinger.

Visscher, A. J., & Coe, R. (Eds.). (2002). *School improvement through performance feedback*. Lisse: Swets & Zeitlinger.

Visscher, A. J., & Coe, R. (2003). School performance feedback systems: Conceptualisation, analysis, and reflection. *School Effectiveness and School Improvement*, *14*, 321–349.

Wayman, J. C. (2005). Involving teachers in data-driven decision making: Using computer data systems to support teacher inquiry and reflection. *Journal of Education for Students Placed at Risk*, *10*, 295–308.

Wayman, J. C., Cho, V., & Johnston, M. T. (2007). *The data-informed district: A district-wide evaluation of data use in the Natrona county school district*. Austin: The University of Texas.

Wayman, J. C., Jimerson, J. B., & Cho, V. (2012). Organizational considerations in establishing the data-informed district. *School Effectiveness and School Improvement*, *23*, 159–178.

Wayman, J. C., Spikes, D., & Volonnino, M. R. (2013). Implementation of a data initiative in the NCLB era. In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data-based decision making in education: Challenges and opportunities* (pp. 135–153). Dordrecht: Springer.

Wayman, J. C., & Stringfield, S. (2006). Data use for school improvement: School practices and research perspectives. *American Journal of Education*, *112*, 463–468.

Wohlstetter, P., Datnow, A., & Park, V. (2008). Creating a system for data-driven decision-making: Applying the principal-agent framework. *School Effectiveness and School Improvement*, *19*, 239–259.

Yang, M., Goldstein, H., Rath, T., & Hill, N. (1999). The use of assessment data for school improvement purposes. *Oxford Review of Education*, *25*, 469–483.

Zupanc, D., Urank, M., & Bren, M. (2009). Variability analysis for effectiveness and improvement in classrooms and schools in upper secondary education in Slovenia: Assessment of/for Learning Analytic Tool. *School Effectiveness and School Improvement*, *20*, 89–122.

Downloaded by [Universiteit Twente] at 02:33 11 November 2015