Studies in Educational Evaluation 39 (2013) 144–152

Contents lists available at SciVerse ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

Interpretation of the score reports from the Computer Program LOVS by teachers, internal support teachers and principals

Fabienne M. van der Kleij [a,b,*], Theo J.H.M. Eggen [a,b]

[a] Cito, Postbus 1034, 6801 MG Arnhem, The Netherlands
[b] University of Twente & RCEC, University of Twente, Postbus 217, 7500 AE Enschede, The Netherlands

ARTICLE INFO

Article history:
Received 25 September 2012
Received in revised form 10 April 2013
Accepted 14 April 2013

Keywords:
Student evaluation
School-based evaluation
Data feedback
Score report interpretation

ABSTRACT

Data-driven decision making, such as the decision making that is conducted through the use of pupil monitoring systems, has become increasingly popular in the Netherlands, as it is considered to have promise as a means of increasing pupils' learning outcomes. The reports generated by the pupil-monitoring Computer Program LOVS (Cito) provide educators with reliable and objective data feedback; however, research has suggested that many users struggle with interpreting these reports. This study aims to investigate the extent to which the reports are correctly interpreted by educators, and to identify various potential stumbling blocks with regards to the interpretation of the reports. The results suggest that users encounter many stumbling blocks in these reports and often cannot interpret them entirely correctly.

When data about students are used to inform decisions in the school, it is referred to as Data-Driven Decision Making (DDDM). Through DDDM, one can guide education based on the outcomes of measurements in both a diagnostic and evaluative way (Ledoux, Blok, Boogaard, & Krüger, 2009). School performance feedback systems (SPFS) are external party systems that aim to provide schools with insight into the outcomes of the education they have provided (Visscher & Coe, 2002). SPFS provides schools with feedback on a systematic basis (Fitz-Gibbon & Tymms, 2002). Ultimately, this feedback aims to improve the quality of education within the school (Verhaeghe, 2011). Pupil-monitoring systems are a kind of SPFS that have been developed primarily to monitor the individual progress of pupils. Pupil monitoring systems are important in DDDM, since the data about learning progress at the pupil level form an important source of information for decisions at all levels of the school.

The Dutch Ministry of Education Culture and Science (2010) promotes DDDM. The Ministry distinguishes four levels at which DDDM can be aimed: the school board level, the school level, the class level and the level of the individual pupil. For the successful implementation of DDDM, the Ministry uses five indicators:

- the annual evaluation of the learning outcomes of pupils;
- the frequent evaluation of the educational process;
- the systematic monitoring of pupils' progress by teachers;
- the quality of the testing system; and
- the evaluation of the effects of interventions.

The indicators point out that the ministry strives towards a schoolwide implementation of DDDM. The Dutch DDDM policy requires the entire school team to evaluate the education based on test results. Principals are expected to conduct schoolwide evaluations for both internal (school improvement – formative) and external (accountability – summative) purposes. The ministry (2010) expects teachers to systematically monitor their pupils' progress, meaning that they have insight into the capacities, potentials and limitations of their pupils based on the results of a pupil monitoring system and classroom assessment. Internal support teachers are expected to collaborate with the class teachers and to support them in interpreting test results, analysing test results and seeking suitable solutions to learning problems.

DDDM encompasses a systematic and cyclic process. Bennett (2011) has described the cyclic process of educational measurement as consisting of four activities: "...designing opportunities to gather evidence, collecting evidence, interpreting it, and acting on interpretations" (p. 16). This study focuses on the interpretation of test results from Cito's[1] pupil monitoring system for primary education (LOVS).

* Corresponding author at: Cito, Postbus 1034, 6801 MG Arnhem, The Netherlands. Tel.: +31 0263521599.
E-mail addresses: fabienne.vanderkleij@cito.nl (F.M. van der Kleij), theo.eggen@cito.nl (Theo J.H.M. Eggen).

[1] The Institute for Educational Measurement in the Netherlands.

0191-491X/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.stueduc.2013.04.002

The LOVS program encompasses various tests (e.g. Math, reading comprehension and spelling) that can be used to systematically map pupils' learning progress. LOVS tests are primarily meant to provide teachers with insight into the outcomes of the education that has been offered. These insights can subsequently be used to adapt teaching where needed. Approximately 90% of Dutch primary schools use the LOVS tests. The Computer Program LOVS allows the user to process test results and automatically generate pupil reports, group overviews and school reports. In this process, accurate interpretation of the results is of the utmost importance.

Meijer, Ledoux and Elshof (2011) recently published a report about the usability of various pupil monitoring systems in Dutch primary education. The results of this study suggest that users of the Computer Program LOVS have difficulty interpreting the test results, which sometimes results in users making incorrect decisions. In addition, use of the test results by teachers appears to be limited, as interpretation and analysis of the results is mainly executed by internal support teachers. This conclusion is also supported by Ledoux et al. (2009), who claim that teachers are not always involved in the interpretation phase. In addition, multiple studies (Ledoux et al., 2009; Meijer et al., 2011) suggest that the many possibilities offered by the Computer Program LOVS are only used to a limited extent. For example, the trend analyses often remain unused. Various studies from outside the Netherlands have suggested that school staff currently lack the knowledge and skills that are needed to use data to improve the quality of education (Earl & Fullan, 2003; Kerr, Marsch, Ikemoio, Darilek, & Barney, 2006; Ledoux et al., 2009; Meijer et al., 2011; Saunders, 2000; Van Petegem & Vanhoof, 2004; Williams & Coles, 2007; Zupanc, Urank, & Bren, 2009). Vanhoof, Verhaeghe, Verhaeghe, Valcke, and Van Petegem (2011) emphasise that there is little knowledge about the degree to which users are capable of correctly interpreting and analysing data from SPFS; this is a crucial precondition for DDDM.

Moreover, various studies have suggested that a certain degree of 'assessment literacy' is a precondition for a correct interpretation of test results (Earl & Fullan, 2003; Vanhoof et al., 2011; Verhaeghe, 2011). "Assessment literacy refers to the capacity of teachers – alone and together – (a) to examine and accurately understand student work and performance data, and correspondingly, (b) to develop classroom, and school plans to alter conditions necessary to achieve better results" (Fullan & Watson, 2000, p. 457). As data interpretation is necessary for adequately altering conditions to meet pupils' needs, it touches upon one of the basic skills that compromise assessment literacy. Hattie and Brown (2008) noted that when assessment results are displayed graphically, the need for teachers to have a high degree of assessment literacy is reduced because they can make use of their intuition to interpret the assessment results (a). However, they emphasised that teachers do need to be very skilled in transforming their interpretations into meaningful actions for teaching that meet the needs of the learners (b). Mandinach and Jackson (2012) call this 'pedagogic data literacy'. The Computer Program LOVS provides both numerical information in the form of a table and graphical representations, which allows for intuitive interpretations and provides numerical data for further analysis and comparison to instructional goals. However, it is not clear which (basic) level of assessment literacy can be expected of the current teacher population in the Netherlands. Popham (2009) has noted that currently in most pre-service teacher education programs in the United States, courses on educational assessment are not part of the curriculum and no formal requirements exist. This situation is no different in the Netherlands, although the recent developments in the area of DDDM have boosted professional development initiatives.

LOVS is known as a pupil monitoring system that uses advanced psychometric techniques, which results in reliable and valid outcomes about pupil ability. However, whenever users draw incorrect inferences, the validity of the test scores is negatively affected. Being able to correctly interpret pupils' test results is a precondition for the optimal use of the Computer Program LOVS. Besides the above – mentioned lack of knowledge among school staff, it has been suggested that many teachers are uncertain about their own ability to use data for quality improvement (e.g. Earl & Fullan, 2003; Williams & Coles, 2007). On the one hand, there is much to be gained through professional development in regards to the interpretation and use of data feedback. For example, a study by Ward, Hattie, and Brown (2003) pointed out that professional development increased correctness in the interpretation of reports belonging to a pupil monitoring system and also increased communication about test results with colleagues, enhanced user confidence and increased use of the various reports. On the other hand, clear score reports can support users in making correct interpretations (Hattie, 2009; Ryan, 2006; Zenisky & Hambleton, 2012). For example, Hattie and Brown (2008) evaluated whether users off asTTle reports could correctly interpret these reports. The initial 60% that was correct was not found to be satisfactory. The researchers subsequently adjusted features of the reports whereupon the percentage correct increased to over 90%.

In the literature, remarkably little attention is paid to the way users (mis)interpret the score reports. For example, *The Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) contain only a few general standards about score reporting. The possible incorrect or incomplete interpretation of assessment results is an underexposed but important aspect of formative testing (Bennett, 2011). There is scarce research into the characteristics of feedback reports and the effectiveness of various methods used for communicating feedback to users (Verhaeghe, 2011). This is problematic, since feedback reports often contain complex graphical representations and statistical concepts, while users often do not possess statistical skills (Earl & Fullan, 2003; Kerr et al., 2006; Saunders, 2000; Williams & Coles, 2007).

Reports can serve two purposes (Ryan, 2006). First, they can be instructive by informing the target group about pupils' learning progress and the effectiveness of instruction. Second, reports can be used to ensure accountability. This study focuses on their instructive purposes. LOVS primarily aims at informing schools about their own functioning. Recent research, however, suggests that the instructive use of LOVS reports is limited, and teachers struggle with interpreting these reports (Meijer et al., 2011). Most notably, various recent studies suggest that members of the school board (e.g. school principals) have a more positive attitude towards SPFS than teachers (Vanhoof, Van Petegem, & De Maeyer, 2009; Verhaeghe, Vanhoof, Valcke, & Van Petegem, 2011; Zupanc et al., 2009). Zenisky and Hambleton (2012) have recently emphasised that although the body of literature on effective score reporting is growing, investigations of actual understanding among users is needed. This is also needed as part of ongoing maintenance for reports that have already been developed or used for a while. Although the body of research on the interpretation of results from the Computer Program LOVS is growing, user interpretation has not yet been systematically investigated among various user groups. Thus, actually testing users' interpretations and discussing the aspects of the reports could provide insight into whether or not specific features of the score reports cause educators to struggle, in which case, appropriate adaptations can be made. Given the fact that the contents of the score reports can be directly manipulated by the test developers, it seemed appropriate to conduct an

empirical study in order to investigate whether the score reports from the Computer Program LOVS could be improved.

The purpose of this study is to (a) investigate the extent to which the reports from the Computer Program LOVS are correctly interpreted by school principals, internal support teachers and teachers and (b) identify stumbling blocks for teachers, internal support teachers, and principals when interpreting reports from the Computer Program LOVS. Furthermore, the study aims to explore the possible influences of various variables that seem relevant given the literature (e.g. Earl & Fullan, 2003; Meijer et al., 2011; Vanhoof et al., 2009). These variables are training in the use of the Computer Program LOVS (Ward et al., 2003), the number of years of experience using the Computer Program LOVS (Meijer et al., 2011), the degree to which the information from the Computer Program LOVS is perceived as useful (Vanhoof et al., 2009; Verhaeghe et al., 2011; Zupanc et al., 2009), and users' estimates of their own ability to use quantitative test data (Earl & Fullan, 2003; Williams & Coles, 2007).

Theoretical framework

The use of data feedback

The test results from pupil monitoring systems provide users with feedback about pupil performance. This is called data feedback. This feedback is intended to close the gap between a pupil's current performance and the intended learning outcomes (Hattie & Timperley, 2007). Various studies suggest that the actual use of feedback about pupil performance within the school is limited. A possible explanation for the lack of feedback use can be found in the characteristics of the SPFS (Earl & Fullan, 2003; Schildkamp & Kuiper, 2010; Schildkamp & Visscher, 2009; Verhaeghe, Vanhoof, Valcke, & Van Petegem, 2010; Visscher & Coe, 2002). More specifically, in the Dutch context, it can be concluded that the use of data feedback by teachers in primary education is limited (Ledoux et al., 2009; Meijer et al., 2011), although research has suggested that Dutch schools possess sufficient data feedback (Ministry of Education, Culture, and Science (2010). Visscher (2002) has identified several factors that influence the use of data feedback within schools: the design process and characteristics of the SPFS, characteristics of the feedback report and the implementation process and organisational features of the school. This study focuses on the characteristics of the feedback report.

With regard to the use of data feedback from pupil monitoring systems, various types of uses can be distinguished. A distinction can be made between the instrumental use and the conceptual use of the test results (Rossi, Freeman, & Lipsey, 1999; Weiss, 1998). The instrumental use compromises the direct use of findings to take actions were needed. The major form of instrumental use of data feedback from pupil monitoring systems is the instructional use. The conceptual use encompasses the impact test results can have on the way educators think about certain issues. Visscher (2001) distinguishes an additional type of data use, namely the strategic use of data feedback. This type of use includes all sorts of unintended uses of data feedback for strategic purposes, such as teaching to the test or letting certain pupils not sit the test. A correct interpretation of data feedback is especially necessary for adequate instrumental use.

The literature reports several preconditions that have to be met in order for a score report to be used. The contents of the feedback reports should be perceived as relevant, useful and non – threatening (Schildkamp & Teddlie, 2008; Van Petegem & Vanhoof, 2007; Visscher, 2002). Furthermore, the feedback must be reliable, valid and delivered in a timely manner (Schildkamp & Teddlie, 2008; Visscher, 2002; Visscher & Coe, 2003). Moreover, Vanhoof et al. (2011) suggest that the confidence of users in their own ability to use data feedback from a SPFS, and their attitude towards feedback, positively affect the degree to which users are willing to invest in the use of data feedback.

The interpretation of data feedback

The literature distinguishes between data and information (Davenport & Prusak, 1998; Mandinach & Jackson, 2012). Data are objective facts that do not carry meaning. By interpreting data, these facts can be transformed into information – for example, by summarising and computing (Davenport & Prusak, 1998). Subsequently, information can be turned into usable knowledge, which is the basis for a decision about an action. The impact of the action is evaluated using new data; this way, a feedback loop is created (Mandinach & Jackson, 2012). Clear score reports can support users in making correct interpretations (Hattie, 2009; Ryan, 2006; Zenisky & Hambleton, 2012).

Although the literature about score report interpretation and/or misinterpretation is scarce, supporting users in interpreting the reports has recently been addressed as an important aspect of validity (Hattie, 2009; Ryan, 2006). This is especially relevant when test results inform important decisions. An incorrect interpretation can lead to inadequate decisions and, subsequently, inadequate actions. In education, this could mean that learning deficits are not signalled, whereupon the pupil does not get the needed support or additional instruction. In addition, it could mean that weak spots in the effects of instruction are not identified. In other words, whenever the test results are interpreted incorrectly, instruction cannot be tailored to the needs of the pupils. Various researchers have recently highlighted the lack of research about the interpretation of score reports (Hattie, 2009; Ryan, 2006; Verhaeghe, 2011; Zenisky & Hambleton, 2012). In addition, the crucial role of test developers in supporting correct interpretations through clear score reports as an aspect of validity has been emphasised (Hambleton & Slater, 1997; Hattie, 2009; Ryan, 2006; Zenisky & Hambleton, 2012). Ryan has emphasised the need to take into account the characteristics of target groups, because, for example, not all users are equally able to interpret statistical data.

Standards for score reports

The standards for score reports described in *The Standards for Educational and Psychological Testing* (AERA et al., 1999) are of a general nature. These guidelines are specifically targeted at validity issues; validity is described as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). They comprise nine standards that apply to score reports. From these standards, it can be concluded that the test developer has a certain amount of responsibility for valid interpretation and use of the test results. This responsibility is shared with the target group to which the report communicates (Ryan, 2006). Hattie (2009) argues that, recently, the user has increasingly been held responsible for a correct interpretation of the test results. He advocates that test developers should pay more attention to the design of their reports. According to Hattie, this is necessary in order to make sure that the users interpret the test results as the test developer intended and then draw adequate inferences and undertake responsible actions.

Method

Exploration and scope refinement

In order to explore the problem, a group of experts was consulted. These experts comprised educational advisers, trainers,

and researchers who often come into contact with users of the Computer Program LOVS. The experts were asked which (aspects of the) reports caused users to struggle. The experts were approached through e-mail, and their responses were discussed in face-to-face meetings and/or in telephone conversations. Furthermore, a researcher attended two training sessions with educational professionals in order to gain insight into the nature of the problem. From this exploration, five reports generated by the Computer Program LOVS were selected for the study: the pupil report, the group overview (one test-taking moment), ability growth, trend analysis, and alternative pupil report. These five reports were chosen based on the frequency with which they have been used within schools and the degree to which the reports are interpreted incorrectly (based on the experts' experience).

In this study, data about user interpretations were collected using multiple methods. Focus groups were formed at two different schools. These groups consisted of teachers, internal support teachers, school principals and other school personnel. Furthermore, the interpretation ability of a group of users was measured using a questionnaire. A multi-method design was chosen for multiple reasons. First, the data from the focus group meetings were used to validate the plausibility of the answering options in the questionnaire. Thus, the qualitative data helped to develop the quantitative instrument. Furthermore, qualitative data from the focus group meetings could lead to in-depth insights into the results found in the questionnaire data with respect to why certain aspects of the reports may be interpreted incorrectly and what possible solutions could be applied to these misinterpretations.

After the second of two rounds of consultations with the experts, the underlying skills necessary for interpreting the score reports were chosen and then mapped into a test specification grid. With regard to knowledge, the following aspects were distinguished:

- knowing the meaning of the level indicators (A–E and I–V);
- knowing the position of the national average within the different levels;
- knowing the meaning of the score interval around the ability; and
- knowing that the norms for groups differ from those for individual pupils with regard to the level indicators.

With respect to interpretation, the following aspects were distinguished:

- judging growth based on ability and signalling negative growth;
- understanding to which level a trend is referring;
- interpreting ability growth as opposed to ability scores;
- understanding whether the growth of the group is under or above the national average;
- comparing the level at which the pupil is functioning to the grade the pupil is in; and
- understanding when a level correction has taken place.

The test grid was used to aid the systematic analysis of the qualitative data from the focus group meetings, and served as a basis for the questionnaire development.

Focus groups

Measurement instruments and procedure

Through focus group meetings at two schools, qualitative data were gathered about the interpretation process and the possible misinterpretations. The focus groups were set up in the form of a group discussion (Newby, 2010). The focus group meetings took place at the participating schools. An educational adviser fulfilled the role of moderator and led the discussion while one of the researchers took notes. The moderator explained the motivation for conducting the study and the purpose of the study. Next, a general investigation of user experience is followed. The moderator asked the participants the following questions: "What are your experiences with the Computer Program LOVS?" "How are the results being used?" "How experienced are you in the use of the Computer Program LOVS?" Subsequently, the participants were shown displays that showed screenshots of the reports (identical to the ones used in the questionnaire), which served as the main stimuli. The use of standardised displays has the benefit that the main stimuli were identical for all participants (Newby, 2010). For each report, approximately 10 min were spent discussing its content. With each report, the moderator asked at least three questions: "What do you see?" "What do you think are striking features of this report?" "What would you conclude from this report?" The meetings at both schools took approximately one and a half hours. The researcher wrote reports on the meetings, which were sent to the contact person in each school for verification (member checking, Creswell & Plano Clark, 2007).

Respondents

The focus group at School 1 consisted of four teachers and two school principals, all female. Both school principals had approximately two years' experience in the role of internal support teacher before they became principal. Five of the six participants had five or more years' experience using the Computer Program LOVS; one teacher had worked with it for over a year. All of the teachers were currently teaching in the lower grades.

The focus group at School 2 consisted of a female teacher, a female adjunct school principal, a female internal support teacher, and a male ICT teacher/coordinator. The participants had five to ten years' experience using the Computer Program LOVS. The internal support teacher has been in this function for four years. The teacher works in grade six and is also coordinator of the upper grades.

Data analysis

The participants' responses to the three questions posed with each report were summarised. Also, other relevant responses as a result of further discussion were listed. Subsequently, users' responses were systematically mapped onto the test grid. This analysis allowed the researchers to see which stumbling blocks appeared to be present in relation to the required knowledge and skills for the various reports, along with the users' suggestions for improvement. Furthermore, aspects that led to confusion that did not relate directly to a specific type of knowledge or skill were listed.

Questionnaire

Measurement instruments and procedure

In order to measure the interpretation ability of the respondents, a questionnaire was constructed in collaboration with the experts. The test grid was used as a basis for constructing the questionnaire in order to come to a representative set of items for measuring the interpretation ability on the selected reports. The plausibility of the alternatives in the questionnaire was evaluated by consulting experts and by analyzing the results of the focus group meetings.

The questionnaire that was used in this study contains 30 items, of which 29 items have a closed-answer format, and one item has an open-answer format. The item with the open-answer format was an item in which respondents could leave remarks and suggestions.

The questionnaire contains nine items about the respondents' background characteristics. The respondents were asked questions

about the following: their gender, the name of their school, their function within the school, which grade they currently teach, their years of experience teaching primary education, what they consider to be their own ability in using quantitative test data as a measure for assessment literacy (Vanhoof et al., 2011), their experience using the Computer Program LOVS and the degree to which they find the information from the reports generated by the Computer Program LOVS to be useful (Vanhoof et al., 2011).

The questionnaire contains twenty items that measure interpretation ability ($\alpha = 0.91$). Of these items, five were intended to measure knowledge and fifteen were intended to measure understanding and interpretation. All items were related to a visual representation of a report. In total, seven visual representations with accompanying items were presented. (Two representations of the pupil report and the group report were provided. The first measured knowledge; the second measured interpretation.) Two to four items were subjected to the respondent about each report. The greater part of the items ($n = 12$) had a multiple response format, which means the respondent could provide multiple answers. The remaining items had a multiple-choice format ($n = 8$), meaning that respondents could only select one answer. The number of options with each item varied from three to six. Participants were granted one point per correct answer, which is the most reliable manner for scoring multiple response items (Eggen & Lampe, 2011). The maximum score on the total questionnaire was 34.

Given that respondents make decisions based on the report, it is of critical importance that they interpret these reports in the correct manner. Therefore, in consultation with the experts, a standard was set. It was expected that the users should be able to answer at least 85% of the items correctly. This corresponds with a score of 29 on the questionnaire.

Respondents

For the questionnaire, two samples were drawn from the customer base of the Computer Program LOVS. The first sample was a random sample consisting of 774 schools. The schools all received a letter requesting them to participate in the study. Schools could send an e-mail if they wanted to participate with one or more staff members. Data were gathered from teachers, internal support teachers, remedial teachers, and school principals. In total, 29 schools signed up for participation in the study (3.7%). Given the large number of non-responses, the researchers decided to draw a second sample. This sample was not random; it consisted of schools that were not selected for participation in a pre-test of one of the LVS tests. The second sample contained 617 schools of which 27 agreed to participate (4.4%).

The questionnaire was filled out online by the respondents. Schools that agreed to participate in the study received an e-mail with a link to the questionnaire, which was distributed within the school by the contact person. In total, nearly 100 respondents from 56 schools filled out the questionnaire (15 males, 81 females, one gender unknown). The relatively large amount of females in the sample is typical for the Dutch primary school teacher population. A recent publication of the Dutch Ministry of Education, Culture, and Science (2011) indicates that, currently, 81% of the teachers in primary education are female. The group of respondents consisted of class teachers (including teachers with an additional task, such as ICT coordinator) ($n = 37$), internal support teachers (including remedial teachers) ($n = 43$), and school principals (including adjunct principals and location managers) ($n = 17$).

Data analysis

The data that were gathered using the questionnaire were analysed both qualitatively and quantitatively. The quantitative analysis was conducted using Classical Tests Theory (CTT) in TiaPlus (2010). The extent to which the reports from the Computer Program LOVS were correctly interpreted was examined using descriptive statistics. Interpretations of various user groups were compared to the standard of 85% correct. Furthermore, the differences between the various user groups were analysed using ANOVA. The relationship with other variables was examined using ANOVA and Pearson correlation analyses. The qualitative analysis was intended to interpret the quantitative data in terms of points of struggle for the various respondent groups on the various reports. For example, whether there were differences between the various user groups with respect to the particular reports was explored.

Results

Focus groups

The results of the focus group meetings suggest that several aspects of the reports caused confusion or a faulty interpretation. For example, in multiple reports, a triangle that points up or down was used. The participants noted that this symbol suggested a particular meaning, namely 'increase or decrease'. However, the symbols were merely meant to indicate grades/groups of pupils or a point in a graph. Furthermore, the use of colour was not always straightforward. For example, in the trend analysis, the colour red carried the meaning 'below average', while green meant 'above average'. In this same report, however, the colour green was also used to indicate groups. Participants noted that this led to confusion. In addition, the use of colour was not always sufficiently distinctive. For example, participants noted that the lines indicating the group average and the national average in the ability growth report were hard to distinguish from one another. Furthermore, the participants noted that the distinction between individual and group norms was not clear. The concept of score interval (90% confidence interval around the ability) was also not clear to most participants. Moreover, none of the participants indicated that they used the score interval in daily practice. Additionally, participants noted that the indications of the axes in the graphs were not always complete and clear.

Questionnaire

The extent to which the reports from the Computer Program LOVS are correctly interpreted. On average, the respondents ($N = 97$) gained a score of 21 on the questionnaire (SD = 8.15), which corresponds with an average percentage correct of 61.76%. This number is well below the standard that was set, namely a score of 29 or 85% correct. Only 13 respondents gained a score of 29 or higher (29.89%); 10 of these were internal support teachers and three were principals. This means that of the internal support teachers, 23.26% realised the expected minimum score. Of the principals, 17.65% reached the expected minimum score. The expected minimum score of 29 was not realised by any of the respondent teachers. The highest score was 28 ($n = 2$).

Interpretations by various user groups. In Table 1, the scores gained by the various groups of respondents are displayed. The score is used as the dependent variable in the analyses as a measure of interpretation ability.

Table 1 shows that there is a considerable amount of variation between the total score of teachers, internal support teachers and principals. The results show that the average score for teachers ($n = 37$) was 17.78. This suggests that of all user groups, teachers struggle most in interpreting the reports of the Computer Program LOVS. The differences between the total scores of the various groups were analysed using ANOVA. The results suggest that there

Table 1
Total score and percentage correct per group.

Group	*n*	Total score max = 34	*SD*	Percentage correct
Teacher	37	17.78	8.86	52.29
Internal support teacher	43	23.95	6.53	70.44
School principal	17	20.53	7.88	60.38

is a significant difference among the groups: $F(2,94) = 6.38$, $p = .003$. Post hoc analysis using the Bonferroni method shows that the total scores of teachers were significantly lower than those of internal support teachers (average difference = –6.17, $p = .002$). The differences between the scores of teachers and school principals (average difference = −2.75, $p = .685$) and the scores of internal support teachers and school principals (average difference = 3.42, $p = .376$) were not significant. These results suggest that teachers are significantly less able to interpret the reports generated by the Computer Program LOVS than internal support teachers.

Identifying stumbling blocks. The results of the questionnaire suggest that there are points of struggle in all five reports, as indicated by the respondents' interpretations. Not all respondents possess the necessary basic knowledge to interpret the reports correctly. For example, the meaning of the level indicators A–E and I–V was not known by all respondents. In addition, not all respondents knew the position of the national average within the different levels. The results suggest that approximately one-quarter of the respondents knew what the score interval means. Furthermore, it appeared to be unclear to respondents why norms for groups deviate from the norms for individual pupils.

With regard to the group reports, respondents mostly struggled with interpreting ability growth as opposed to ability and with signalling negative ability growth. Ability growth was often interpreted as ability.

With regard to the reports at the pupil level, respondents mostly struggled with interpreting ability growth as opposed to ability, understanding when a level correction has taken place, and judging growth using ability. When judging growth, strikingly few people used the score interval.

Next, it was explored whether there were differences between the various user groups with respect to the particular reports. Fig. 1 shows the average proportion correct (*P*'-value) for each item belonging to a certain report, plotted for each user group.

On the *x*-axis, the numbers of the items as they appeared in the questionnaire are depicted. For a clear communication of the results, the items have been ordered based on the report to which they belong. The pupil report and alternative pupil report compromise the level of the pupil, the group reports, ability growth, and trend analysis compromise the level of the group. Items 2, 3, 5, 6 and 14 measure knowledge; the other items measure understanding and interpretation.

For the majority of the items, the *P*-values of the various user groups are well below the standard of 85%. The pattern consistently suggests that internal support teachers are more able to interpret the reports than teachers. However, it must be noted that due to the small sample size, the confidence intervals around the *P*-values are large; therefore, no significant differences are present among various user groups at the item level.

From Fig. 1 it can be seen that Item 2 was the easiest item for all user groups. This item measured knowledge with respect to the meaning of the level indicators A–E, more specifically the meaning of level C with respect to the national average. Striking in this respect are the relatively low average *P*-values for Item 5, which measured knowledge with respect to the meaning of the level indicator A. The average *P*-value among teachers is particularly low. Furthermore, the *P*-values suggest that the users are more knowledgeable about the meaning of the level indicators A–E (e.g. Items 2 and 5) than the level indicators I–V (e.g. Item 14). The hardest item was Item 3, which measured users' knowledge of what the score interval means. Furthermore, Item 10 stands out since both internal support teachers and principals scored on average above the standard, but teachers did not. This item measured the interpretation of ability growth as opposed to ability.

Furthermore, the relationship between various background variables and the interpretation ability was explored.

First, we determined whether there were differences among the three groups in terms of the number of respondents who received training. The differences appeared to be large and significant, $F(3,94) = 19.38$, $p < 0.001$. In the group of teachers, only 5% indicated that they had received some kind of training in the use of the Computer Program LOVS in the last five years. In the group of internal support teachers, 42% had received training, and the majority of the school principals, namely 77%, had received training. Whether or not a respondent had received training in the use of the Computer Program LOVS did not appear to be significantly related to the total score, $F(1,95) = 0.71$, $p = 0.403$.

The number of years' experience using the Computer Program LOVS did not relate to interpretation ability (0–5, 6–10, >10 years) ($F(2,94) = 1.11$, $p = 0.331$).

Furthermore, we examined whether the degree to which the information generated by the Computer Program LOVS is perceived as useful relates to interpretation ability. No evidence was found for such a relationship ($F(2, 93) = 1.51$, $p = .227$). The respondents indicated that they perceived the information generated by the Computer Program LOVS to be a little bit useful ($n = 5$), useful ($n = 37$) or very useful ($n = 54$). For the degree in which information generated by the Computer Program LOVS is perceived to be useful as the dependent variable, the ANOVA results, with function as a factor, suggest that there is a significant difference between respondents in various functions: $F(3,93) = 4.82$, $p = 0.01$. Post hoc analysis indicates that the degree to which the information generated by the Computer Program LOVS is perceived as useful differs significantly between teachers and internal support teachers (average difference = −0.35, $p = 0.025$), and between teachers and school principals (average difference = −0.43, $p = 0.039$). Thus, internal support teachers and school principals were more positive than teachers with regards to the usefulness of information generated by the Computer Program LOVS.

Next, in order to investigate the relationship between the respondents' estimates of their own ability in using quantitative test data and their measured ability, a two-sided Pearson correlation analysis was conducted. The results suggest a moderately positive relationship, which is significant: $r(95) = 0.25$, $p = 0.013$. None of the respondents estimated their own ability in using quantitative test data as 'not at all able' or 'not able'. Of the respondents, 15% judged themselves as 'a little bit able' (0), 64% judged themselves as 'able' (1), and 18% judged themselves as 'very able' (2). School principals had the highest estimation of their own ability and teachers the lowest. The estimation of their own ability differed significantly between respondents in various functions, $F(3,94) = 11.64$, $p < 0.001$. The results of post hoc analyses indicate that teachers estimate their own ability to be significantly lower than internal support teachers (average difference = –0.51, $p < 0.001$) and school principals (average difference = −0.59, $p = 0.001$). On average, teachers judged themselves just above 'a little bit able' ($M = 0.7$, SD = 0.57). Thus, teachers judged themselves to be just above 'a little bit able' in using quantitative test data, and none of the teachers judged themselves to be 'not at all able' or 'not able'. However, it must be noted that teachers judged their own ability at a significantly lower level than respondents in a different function.

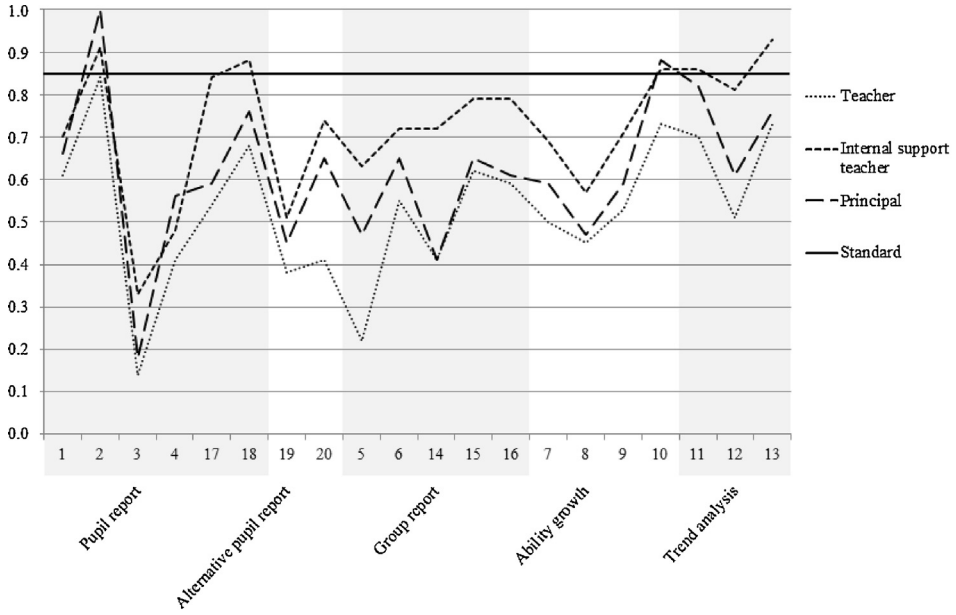

Fig. 1. Average proportion correct for the three user groups at the item level.

Conclusion and discussion

This study explored the extent to which the reports from the Computer Program LOVS are correctly interpreted by school principals, internal support teachers and teachers. Furthermore, the study attempted to identify possible stumbling blocks concerning the interpretation of the score reports in the Computer Program LOVS. By conducting two focus group meetings and administering a questionnaire, both qualitative and quantitative data were gathered. In the quantitative analyses, distinctions were made among teachers, internal support teachers (including remedial teachers) and school principals.

Results from previous studies (e.g. Meijer et al., 2011) have suggested that users of the Computer Program LOVS do not interpret the reports generated by the Computer Program LOVS completely correctly. The results suggest that users have many stumbling blocks in the current reports generated by the Computer Program LOVS. Teachers seem to experience difficulties in interpreting both the reports at the group level and at the pupil level.

The results of the questionnaire suggest that teachers, internal support teachers, and principals have problems with interpreting all five reports. Less than 30% of the respondents scored at or above the standard of 85% correct. Moreover, the results suggest that not all users have the basic knowledge that is required to correctly interpret the reports. For example, the meaning of the levels A–E and I–V and the meaning of the score interval were not well understood, except for the meaning of the level C. There were significant differences among the various respondent groups in terms of the total scores on the questionnaire. The total scores of teachers were significantly lower than those of internal support teachers. The difference between the scores of teachers and school principals was not significant. When looking at the results at the item level, the pattern consistently suggests that internal support teachers are most able when it comes to interpreting the reports.

A major question of this study related to identifying stumbling blocks for users in the interpretation of reports generated by the Computer Program LOVS. The results of the questionnaire suggest that with regard to the reports at the group level, respondents mostly struggled with interpreting growth in ability as opposed to interpreting ability and signalling negative ability growth. The growth in ability was often interpreted as the ability level. With respect to the reports at the pupil level, respondents mostly struggled with the interpretation of growth in ability as opposed to ability, understanding when a level correction has taken place, and the interpretation of growth in ability. When interpreting growth in ability, strikingly few people used the score interval. The results of the focus group meetings are fairly consistent with the results found in the questionnaire with respect to the stumbling blocks in the interpretation of the reports. The results suggest that a number of aspects within the reports caused confusion or faulty interpretations. For example, the use of symbols and colours was not always clear and unambiguous. It also appeared that the indications of the axes in the graphs were not always complete. The concept of score interval appeared to be difficult for focus group participants to understand. Not surprisingly, the score interval was not used in practice by focus group participants. Previous research (Hambleton & Slater, 1997; Zenisky & Hambleton, 2012) on the interpretation of score reports already indicated that statistical concepts related to confidence levels are often ignored by users of the reports because users do not find them meaningful. There appears to be a conflict between the standards for score reports (AERA et al., 1999), which prescribe that confidence levels should be reported, and the data literacy of those who are used to score reports. One could question the usefulness of reporting confidence levels when they are neither understood nor used according to the test developer's intention.

In this study, the possible influences of various variables were explored. Whether or not a respondent had received training in the use of the Computer Program LOVS appeared not to be related to their interpretation ability. However, we did find a substantial and significant difference between the three groups with regard to having received training in the use of the Computer Program LOVS. Strikingly, only 5% of the teachers had received training. This is alarming given that the entire school team is expected to evaluate the education based on test results (Ministry of Culture, Education and Science, 2010) and the limited attention that is currently paid to assessment literacy in teacher pre-service programs. Neither was a relationship found between the number of years of experience using the Computer Program LOVS and interpretation ability. However, in order to make substantial claims about the effects of training and experience, additional research is needed. In this study, for example, which training the respondent had followed was not measured nor was the duration or intensity of

this training. However, various researchers have emphasised the need for good support with regard to the use of data feedback in schools (Schildkamp & Teddlie, 2008; Schildkamp, Visscher, & Luyten, 2009; Van Petegem & Vanhoof, 2007; Verhaeghe et al., 2010; Visscher & Coe, 2003; Zupanc et al., 2009). It would be worthwhile to study the effects of professional development on the interpretation and use of data feedback. For example, recent research (Staman, Visscher, & Luyten, 2013) suggests that teachers can benefit much from an intensive schoolwide training programme in DDDM, focusing on, among other things, the interpretation of test results.

Visscher (2002) has emphasised that not only do the characteristics of the feedback and the feedback system determine to what degree feedback will be used, but the perceptions of the users are also important. Moreover, a negative attitude towards performance feedback can be an obstacle for feedback use (Bosker, Branderhorst, & Visscher, 2007). In this study, the degree to which respondents indicated that they perceived the information generated by the Computer Program LOVS to be useful for their own education did not relate to their interpretation ability. The respondents indicated that they perceived the information generated by the Computer Program LOVS as 'a little bit useful', 'useful', or 'very useful'. The difference between the responses from respondents with various functions was significant. Class teachers experienced the information from the Computer Program LOVS as significantly less useful than internal support teachers and school principals. The finding that class teachers experienced the results from the Computer Program LOVS as less useful than school principals is in line with results from previous studies (Vanhoof et al., 2009; Verhaeghe et al., 2011; Zupanc et al., 2009). According to Meijer et al. (2011), the usability of a pupil monitoring system does not only depend upon the characteristics of the system, but also on how users deal with the system. Meijer et al. claim that users of pupil monitoring systems need to become aware that the results provide useful information about the progress of pupils. Ledoux et al. (2009) also suggest that teachers see DDDM more like an additional burden rather than part of their professional responsibilities. Therefore, the researchers suggest that if data-driven practices in the classroom are to be stimulated, teachers should be made aware of the usefulness and value of the results of a pupil monitoring system for their own education.

Various studies have pointed out that many educators are unsure about their own ability to use data for school improvement practices (e.g. Earl & Fullan, 2003; Williams & Coles, 2007). The results from the questionnaire suggest that all the respondents judged themselves to be 'a little bit able', 'able', or 'very able' to deal with quantitative test data. It is striking that none of the respondents judged themselves to be 'not at all able' or 'not able'. Thus, these results contrast with results from previous studies. Class teachers did judge their own ability to be lower than internal support teachers and school principals, but they still think of themselves as 'a little bit able' to handle quantitative test data. Vanhoof et al. (2011) suggest that the degree in which feedback is actually used is affected by the level of confidence SPFS users have in their own knowledge and ability to use data, as well as by their attitude towards feedback. Thus, the results from this study suggest that these preconditions for feedback use have been met. Moreover, respondents appeared to be able to make a good estimate of their own ability in handling quantitative test data.

This study was limited by the size of the sample. Because the sample was limited and not completely randomly drawn, the results of this study can only be generalised to a limited degree. A certain amount of self-selection by the respondents also took place. Because of this, the results are possibly more positive than they normally would be (e.g. with regard to perceived usefulness). For this study, the selection of five reports was made based on the frequency with which they have been used within schools and the degree to which they have been interpreted incorrectly. If the researchers had chosen different reports, this might have led to different results.

A correct interpretation of the score reports is a necessary precondition for the successful completion of all phases of the evaluative cycle. Indeed, a correct interpretation is directly linked to making a justified decision. Nevertheless, whenever a score report is interpreted correctly, this does not guarantee an appropriate use of the test results in terms of making adaptations to the learning process. Moreover, assessment literacy is not limited to the correct interpretation of test results, it also taps into the ability to transform knowledge about what pupils know and can do into meaningful instructional actions (Fullan & Watson, 2000; Mandinach & Jackson, 2012; Popham, 2009). Future research should point out the extent to which users are capable of transforming data feedback into instructional actions.

An important lesson to be learnt is that although the reports from the Computer Program LOVS have been in use for a couple of years, many users struggle with interpreting the reports. The authors follow Zenisky and Hambleton (2012) in their advice that test score reporting should receive considerable attention by test developers even after the initial developmental stage. Thus, test developers should monitor whether the test results are being used as intended.

It seems worthwhile to examine whether redesigned score reports would be interpreted more correctly. Although the researchers acknowledge that the contextual factors (e.g. assessment literacy, time, pressure and support) also impact the extent to which the reports are interpreted correctly, the test developer is primarily responsible for ensuring validity by way of clear score reports (Hattie, 2009; Ryan, 2006; Zenisky & Hambleton, 2012).

Acknowledgements

The authors would like to thank Jacqueline Visser for facilitating this study; Gerben Veerbeek for assisting in the focus group meetings; the experts consulted – Ilse Papenburg, Ilonka Verheij, and Laura Staman – for generously sharing their expertise; Ronalde Engelen and Servaas Frissen for their research assistance; the respondents who kindly filled out the questionnaire; and the participants of the focus groups for sharing their professional expertise and time.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*, 5–25 http://dx.doi.org/10.1080/0969594X.2010.513678.

Bosker, R. J., Branderhorst, E. M., & Visscher, A. J. (2007). Improving the utilization of management information systems in secondary schools. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice, 18*(4), 451–467 http://dx.doi.org/10.1080/09243450701712577.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge. How organizations manage what they know*. Boston: Harvard Business School.

Earl, L., & Fullan, M. (2003). Using data in leadership for learning. *Cambridge Journal of Education, 33*(3), 383–394.

Eggen, T.J.H.M., & Lampe, T. T. M. (2011). Comparison of the reliability of scoring methods of multiple–response items, matching items, and sequencing items. *Cadmo, 19*, 85–104 http://dx.doi.org/10.3280/CAD2011-002008.

Fitz-Gibbon, C. T., & Tymms, P. (2002). Technical and ethical issues in indicator systems: Doing things right and doing wrong things. *Educational Policy Analysis Archives, 10*(6), 1–28, Retrieved from http://epaa.asu.edu/ojs/article/view/285.

Fullan, M., & Watson, N. (2000). School-based management: Reconceptualizing to improve learning outcomes. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice, 11*, 453–473 http://dx.doi.org/10.1076/sesi.11.4.453.3561.

Hambleton, R. K., & Slater, S. C. (1997). Are NEAP executive summary reports understandable to policy makers and educators? CSE Technical Report 430. Los Angeles: National Centre for Research on Evaluation, Standards, and Student Teaching.
Hattie, J. A., & Brown, G. T. L. (2008). Technology for school-based assessment and assessment for learning: Development principles from New Zealand. *Journal of Educational Technology Systems, 36*, 189–201 http://dx.doi.org/10.2190/ET.36.2.g.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112 http://dx.doi.org/10.3102/003465430298487.
Hattie, J. (2009). Visibly learning from reports: The validity of score reports. *Online Educational Research Journal* Retrieved September 15, 2011 from http://www.oerj.org/View?action=viewPDF&paper=6.
Kerr, K. A., Marsch, J. A., Ikemoio, G. S., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvement: Actions, outcomes, and lessons from three urban districts. *American Journal of Education, 112*(4), 403–420.
Ledoux, G., Blok, H., Boogaard, M., & Krüger, M. (2009). *Opbrengstgericht werken; over de waarde van meetgestuurd onderwijs*. [Data-driven decision making; About the value of measurement oriented education]. Amsterdam, the Netherlands: SCO-Kohnstamm Instituut.
Mandinach, E. B., & Jackson, S. S. (2012). *Transforming teaching and learning through data-driven decision making*. Thousand Oaks, CA: Corwin.
Meijer, J., Ledoux, G., & Elshof, D. P. (2011). *Gebruikersvriendelijke leerlingvolgsystemen in het primair onderwijs*. [User-friendly pupil monitoring systems in primary education]. Amsterdam: SCO Kohnstamm Institute.
Ministry of Education, Culture, and Science. (2010). *Opbrengstgericht werken in het basisonderwijs: een onderzoek naar opbrengstgericht werken bij rekenen–wiskunde in het basisonderwijs*. [Data–driven decision making in primary education: A study on data-driven decision making in maths in primary education]. The Hague, the Netherlands: Ministry of Education, Culture, and Science.
Ministry of Education, Culture, and Science. (2011). *Nota werken in het onderwijs 2012*. [Note working in education 2012]. The Hague, the Netherlands: Ministry of Education, Culture, and Science.
Newby, P. (2010). *Research methods for education*. Harlow, England: Longman.
Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice, 48*, 4–11 http://dx.doi.org/10.1080/00405840802577536.
Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (1999). *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage.
Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 677–710). Mahwah, NJ: Lawrence Erlbaum.
Saunders, L. (2000). Understanding schools' use of 'value added' data: The psychology and sociology of numbers. *Research Papers in Education, 15*(3), 241–258.
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 http://dx.doi.org/10.1016/j.tate.2009.06.007.
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 http://dx.doi.org/10.1080/13803610802048874.
Schildkamp, K., & Visscher, A. J. (2009). Factors influencing the utilization of a school self-evaluation instrument. *Studies in Educational Evaluation, 35*, 150–159 http://dx.doi.org/10.1016/j.stueduc.2009.12.001.
Schildkamp, K., Visscher, A., & Luyten, H. (2009). The effects of the use of a school self-evaluation instrument. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice, 20*, 69–88 http://dx.doi.org/10.1080/09243450802605506.
Staman, L., Visscher, A. J., & Luyten, H. (2013). The effects of training school staff for utilizing student monitoring system data. In D. Passey, A. Breiter, & A. J. Visscher (Eds.), *Next generation of information technology in education management* (pp. 3–14). Heidelberg, Germany: Springer.
TiaPlus (Version 2010) [Computer software]. (2010). Arnhem: Cito.
Vanhoof, J., Van Petegem, P., & De Maeyer, S. (2009). Attitudes towards school self-evaluation. *Studies in Educational Evaluation, 35*, 21–28 http://dx.doi.org/10.1016/j.stueduc.2009.01.004.
Van Petegem, P., & Vanhoof, J. (2004). Feedback over schoolprestatieindicatoren als strategisch instrument voor schoolontwikkelingen [Feedback about school performance indicators as a strategic instrument for school development]. *Pedagogische Studiën, 81*, 338–353.
Van Petegem, P., & Vanhoof, J. (2007). Towards a model of effective school feedback: School heads' point of view. *Educational Research and Evaluation, 13*, 311–325 http://dx.doi.org/10.1080/13803610701702522.
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 http://dx.doi.org/10.1080/03055698.2010.482771.
Verhaeghe, G. (2011). School performance feedback systems: Design and implementation issues. Unpublished doctoral dissertation: University of Gent.
Verhaeghe, G., Vanhoof, J., Valcke, M., & Van Petegem, P. (2010). Using school performance feedback: Perceptions of primary school principals. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice, 21*, 167–188 http://dx.doi.org/10.1080/09243450903396005.
Verhaeghe, G., Vanhoof, J., Valcke, M., & Van Petegem, P. (2011). Effecten van ondersteuning bij schoolfeedbackgebruik [Effects of support in school feedback use]. *Pedagogische Studiën, 88*(2), 90–106.
Visscher, A. J. (2001). Public school performance indicators: Problems and recommendations. *Studies in Educational Evaluation, 27*(3), 199–214.
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: An International Journal of Research, Policy and Practice, 14*, 321–349 http://dx.doi.org/10.1076/sesi.14.3.321.15842.
Ward, L., Hattie, J. A. C., & Brown, G.T. (2003). The evaluation of asTTle in schools: The power of professional development. AsTTle technical report 35, University of Auckland/New Zealand Ministry of Education.
Weiss, C. H. (1998). Have we learned anything new about the use of evaluation? *American Journal of Evaluation, 19*(1), 21–33.
Williams, D., & Coles, L. (2007). Teachers' approaches to finding and using research evidence: An information literacy perspective. *Educational Research, 49*, 185–206 http://dx.doi.org/10.1080/00131880701369719.
Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice, 31*, 21–26 http://dx.doi.org/10.1111/j. 1745-3992.2012.00231.x.
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: An International Journal of Research, Policy and Practice, 20*, 89–122 http://dx.doi.org/10.1080/09243450802696695.

Fabienne M. van der Kleij is a Ph.D. candidate at the Research Center for Examinations and Certification (RCEC), a collaboration between Cito and the University of Twente. She is a member of Cito's Psychometric Research Center. Her specialisations are: feedback effectiveness, computer-based assessments, assessment for learning, data-driven decision making and diagnostic testing.

Theo J.H.M. Eggen is a member of Cito's Psychometric Research Center. He has major experience as a consultant in educational measurement at the University of Twente, at Cito, and internationally. He is a full professor of psychometrics at the University of Twente. His specialisations are: Item Response Theory, quality of tests, (inter)national assessment and computerised adaptive testing. He is the scientific director of the RCEC and president of the International association for computerised adaptive testing (IACAT).