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