Computer-Based Feedback in Formative Assessment

Fabienne M. van der Kleij

COMPUTER-BASED FEEDBACK IN FORMATIVE ASSESSMENT

Fabienne Michelle van der Kleij

Graduation Committee

Chairman	Prof. Dr. K.I. Oudenhoven-van der Zee
Promoter	Prof. Dr. Ir. T.J.H.M. Eggen
Members	Prof. Dr. J. Baird
	Prof. Dr. F.J.G. Janssens
	Prof. Dr. W. Kuiper
	Dr. D.M.A. Sluijsmans
	Dr. Ir. B.P. Veldkamp

Van der Kleij, Fabienne Michelle
Computer-based feedback in formative assessment
PhD Thesis University of Twente, Enschede, the Netherlands. – Met een samenvatting in het Nederlands.
ISBN: 978-90-365-3580-9
doi:10.3990/1.9789036535809
Printed by Ipskamp Drukkers, Enschede
Cover designed by Jamila van der Kleij and Fabienne van der Kleij

This research was supported by Cito, Institute for Educational Measurement in the Netherlands.

COMPUTER-BASED FEEDBACK IN FORMATIVE ASSESSMENT

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Thursday, December 19th, 2013 at 14.45

by

Fabienne Michelle van der Kleij
born on March 13th, 1988
in Beverwijk, the Netherlands

This dissertation has been approved by the promoter:

Prof. Dr. Ir. T.J.H.M. Eggen

Contents

Chapter 1. General Introduction

Although researchers have long advocated the potentially positive influences of formative assessment on learning, its implementation has been challenging. One challenge is the alignment of formative and summative practices. Other complicating factors are, for instance, high teacher workload, class size, frequency of assessments and accompanying feedback. Computers could offer a solution to overcoming some of these obstacles. For example, computer-based assessments can automatically be scored, students can be provided with automatically generated feedback, and reports that provide information on student learning can automatically be generated for use by teachers and others. Although this may seem a solution, the extent to which the computer can help facilitate formative assessment has not been systematically investigated to date. Furthermore, recent publications have highlighted the lack of a uniform definition of formative assessment. However, a set definition is necessary for both implementing formative assessment and evaluating its effectiveness.

The conceptual framework presented next first addresses the broader context of this dissertation, which investigates assessment that intends to inform and support learning—formative assessment. Subsequently, feedback, a key aspect of formative assessment, will be elaborated. The last part of the conceptual framework concerns the use of computers in educational assessment. Some aspects of the conceptual framework will be discussed in depth in subsequent chapters, while some are used only to frame the chapters of this dissertation within a broader context of the research.

1.1 Formative Assessment

In education, assessments are frequently used to obtain information about student achievement and classroom processes. A common way to assess student achievement is by a test, which can be described as "an instrument or systematic procedure for observing and describing one or more characteristics of a student using either a numerical scale or a classification scheme" (Brookhart & Nitko, 2008, p. 5). However, not all testing activities are conducted in a planned and systematic manner. The term *assessment* is often used to signify a broader conception of testing as gathering information about student learning. Assessment encompasses the use of a broad spectrum of instruments, such as paper and pencil tests, projects, and observations (Stobart, 2008).

Assessments have different purposes, and the information gathered can be used at different levels of education for decision making, for example, at the level of student, class, school, or country (Brookhart & Nitko, 2008; Wiliam, Kingsbury, & Wise, 2013). This dissertation focuses on assessments that are intended to be used at the level of the student, class, or school. In addition to assessment, the term *evaluation* is used, which refers to the use of assessment data to make decisions concerning the quality of education at a higher aggregation level than the level of the learner or the class (Harlen, 2007; Shepard, 2005).

1.1.1 Summative and Formative Assessment

Often a distinction is made between the summative and formative purposes of assessment. Scriven (1967) introduced the terms summative and formative in the context of program evaluations. Bloom, Hastings, and Madaus (1971) first applied these terms in the context of assessment. Assessment results are summative in purpose when they are intended to play a role in making decisions about the mastery of a defined content domain. For example, these concern decisions regarding selection, classification, certification, and placement (Sanders, 2011). In other words, it summarises the level of performance over a certain defined content domain and time span, such as a course or study programme (Sadler, 1989). However, if assessment results are intended to inform and steer the learning process, assessment is formative in purpose.

Some discussions about the distinctions between formative and summative assessment have focused on the timing of the assessment, that is, whether it occurs during the learning process (formative) or after the learning has taken place (summative). However, numerous researchers have emphasised that a distinction between formative and summative assessment based on time-related characteristics is not useful (Bennett, 2011; Sadler, 1989; Stobart, 2008). Although the distinction between formative and summative assessment indicates within their intended uses, summative and formative assessments are not mutually exclusive in their purposes. Namely, they can coexist as the primary and secondary purposes of the same assessment (Bennett, 2011). Furthermore, there is a difference between the purpose and the function of assessments. The purpose for which the assessment has been designed largely influences the feasibility of the assessment results in serving a particular purpose, although sometimes test results are used for purposes they were not intended to serve (Stobart, 2008). Thus, the way in which assessment results are actually used determines their function, while the goal for which they have been initially designed determines the purpose. For example, some assessments are intended to make pass/fail decisions about students at the end of a course (summative). However, when they are accompanied by qualitative feedback (formative), these same assessments could inform students about the quality of their work and areas they need to improve. Moreover, sometimes assessments can have intentionally different purposes for different target groups at various levels in the education system. For example, assessments that are intended to support learning at the levels of the individual and the class (formative) can also to be used at the school level to monitor the progress of this group of pupils (summative and formative).

1.2.2 Features of Formative Assessment

The effectiveness of formative assessment is widely acknowledged. However, these claims are not always well grounded, which is, amongst other factors, related to the lack of a uniform definition of the concept of formative assessment (Bennett, 2011). Formative assessment is a broad concept that comprises many definitions (e.g., assessment for learning, and diagnostic testing; Bennett, 2011; Johnson & Burdett, 2010). Formative assessment is thus an umbrella term that covers various approaches based on different learning theories (Briggs, Ruiz-Primo, Furtak, Shepard, & Yin, 2012). The main feature that these approaches have in common is that the evidence that has been gathered using assessments is interpreted and can subsequently be used to change the learning environment in order to meet learners' needs (Wiliam, 2011).

A crucial aspect of formative assessment is *feedback* (Bennett, 2011; Brookhart, 2007; Sadler, 1989; Shepard, 2005; Stobart, 2008). In the context of formative assessment, feedback can be defined broadly in two perspectives (Sadler, 1989): 1) the information resulting from assessments that provides teachers and other stakeholders with insights into student learning, which can be used to adapt instruction to the needs of learners; 2) feedback provided to students based on their responses to an assessment task that is intended to steer their learning processes directly.

Numerous researchers have attempted a satisfactory definition of formative assessment (Black & Wiliam, 2009). Furthermore, Brookhart (2007) noted that the definition of formative assessment has expanded over time. Nevertheless, to date a generally accepted definition has not emerged in the literature. In order derive a clear conceptualisation of the term formative assessment in this dissertation, various definitions will be discussed, along with the features of formative assessment that are relevant to this dissertation, with the aim of proposing a definition of the term.

Bloom et al. (1971) introduced the term formative assessment. They described the purpose of this form of assessment as providing students with information or feedback on their learning, thereby providing suggestions for future learning. The popularity of formative assessment increased rapidly after the publication of Black and Wiliam's (1998a, 1998b, 1998c) studies, which highlighted the potentially positive effects of formative assessment on students' learning outcomes. Black and Wiliam defined assessment as formative "... Only when comparison of actual and reference levels yields information which is then used to alter the gap" (1998c, p. 53), where the gap refers to the distance between the actual performance and the goal (Sadler, 1989). Some researchers have conceptualised formative assessment based on its potentially positive effect on self-regulated learning (e.g., Clark, 2012; Nicol & Macfarlane-Dick, 2006), an aspect of formative assessment that was also stressed in early definitions of formative assessment. Namely, an important aspect of formative assessment is making students aware of what quality means, so they can monitor the quality of their own work while they are learning (Sadler, 1989). Sadler later defined formative assessment as "assessment that is specifically intended to provide feedback on performance to improve and accelerate learning" (1998, p. 77).

Although most research on formative assessment has focused primarily on students, the use of assessment results in a formative way by teachers is also an important aspect of formative assessment (Bloom et al., 1971; Stiggins, 2005). Brookhart (2007) provided a definition of formative classroom assessment in which the role of the teacher is central: "Formative classroom assessment gives teachers information for instructional decisions and gives pupils information for improvement" (p. 43). Another example is McManus's (2008) conceptualisation, which describes formative assessment as a process that results in feedback that can be used to make instructional decisions and to support and steer student learning. Here, the crucial roles of both teachers and students are emphasised. Others have focused on the evidence actually being used for decision making regarding the learning process:

> An assessment functions formatively to the extent that evidence about student achievement elicited by the assessment is interpreted, and used to make decisions that are likely to be better, or about founded, then the decisions that would have been taken in the absence of the evidence. (Wiliam, et al., 2013, p. 9)

Black and Wiliam (1998c) argued that even though an assessment can be formative in purpose, it does not necessarily serve a formative function. Although it is reasonable to assume that for formative assessment to have an effect on learning, the evidence must be used, it is also warranted to define as formative assessments that are *intended* to provide information that can be used to steer future learning. This puts the focus on the purpose, instead of the actual use of the test results. Moreover, it is not clear when, how, by whom, and for which purposes the evidence about student learning should be used in order to qualify as formative assessment. The many phases in the evaluative cycle that characterises educational measurement (Bennett, 2011) are often not fully or not correctly completed (Young & Kim, 2010). Interpreting evidence, making decisions, and subsequently acting on these decisions are distinct steps in the evaluative cycle. In many cases, the actor does not reach the stage of actually using the results to take action. Moreover, it is extremely difficult to find out if and how assessment results are actually being used. Hence, a definition in which the actual use is formulated as a precondition of applying the term formative leads to blurred boundaries instead of a clarified definition. Thus, it is useful to make a distinction between formative assessment (in which the formative potentials may or may not be utilised fully), and *effective* formative assessment. The latter indicates that the information gathered is correctly interpreted, from which justified inferences may be drawn and subsequently inform decisions that are the basis of actions that effectively adapt the learning environment to the student's needs.

In addition, many of the discussions regarding the meaning of formative assessment have evolved around whether it is a process or an instrument. Making such an explicit distinction is not very helpful, however, because both the process and instruments used, as well as the thoughtful interplay between the two, are essential in effectively supporting student learning (Bennett, 2011). Nevertheless, although coherence between the process and instruments is necessary for formative purposes, the process is not always independent of the instruments used. Namely, in many computer-based formative assessments, feedback is embedded and is provided directly to the learner without the intervention of the teacher. Thus, in this case, the process is inherent in the instrument although the feedback provided may not be used, but at least this was the developer's intention.

Because a wide array of practices was referred to as formative, and the meaning was not clear, scholars from the UK (Assessment Reform Group [ARG], 1999) posited the term *assessment for learning* (AfL), in contrast to *assessment of learning*. AfL is an approach to assessment that focuses on the quality of the learning process, instead of merely on students' (final) learning outcomes (Stobart, 2008). Although the use of the new terminology was intended to clarify the meaning of formative assessment, the degree to which clarification has taken place is questionable (Bennett, 2011). For example, a recent paper was entitled "What is assessment for learning?" (see Wiliam, 2011).

Furthermore, the term data-driven decision making (DDDM)—also called data-based decision making (DBDM) in the more recent literature—is often associated with formative assessment. DDDM originated in the USA as a direct consequence of the No Child Left Behind (NCLB) Act, which defines improving students' learning outcomes by focussing on results and attaining specified targets (Wayman, Spikes, & Volonnino, 2013). Furthermore, diagnostic testing (DT) is often referred to as formative assessment. DT was initially used to refer students to special education, particularly those diagnosed as unable to participate in mainstream educational settings (Stobart, 2008). In DT, detailed assessment data about a learner's problem solving are collected to explain his or her learning process and learning outcomes (Crisp, 2012; Keeley & Tobey, 2011). In some studies, the terminology of all these approaches to assessment is used interchangeably and often inappropriately. For example, the literature on DDDM tends to cite sources concerning AfL, but not vice versa (e.g., Swan & Mazur, 2011). Although all these approaches can contribute to and support student learning, they are different in their theoretical underpinnings and the levels to which they apply in the school setting.

In this dissertation, formative assessment is viewed as an overarching and broad concept that encompasses various approaches that include specific ways in which assessment can be used to support learning. The following working definition will be used in this dissertation: *Formative assessment is any assessment that provides feedback that is intended to support learning and can be used by teachers and/or students.*

In this definition, the term support relates to both cognitive and metacognitive support, although the focus of this dissertation concerns the cognitive effects of formative assessment. Furthermore, in this definition, the concept of support is not restricted to the student, but also refers to the learning of teachers, although the focus of this dissertation is student learning. Furthermore, the term feedback refers to both the information resulting from assessments that provide teachers and other stakeholders with insights into student learning and the feedback provided to students based on their responses to an assessment task that is intended to steer their learning processes directly (Sadler, 1989).

1.2 Feedback

Feedback is viewed as one of the most powerful means to enhance student learning (Hattie & Gan, 2011; Hattie & Timperley, 2007). Namely, providing students with feedback makes it possible to fill the gap between what students know and what they are supposed to know, whereas teachers need feedback in order to determine where students are in their learning process and to adapt teaching activities to the actual needs of students (Sadler, 1989; Stobart, 2008). This section elaborates the aspects that influence the effectiveness of feedback on learning and offers a typology of item-based feedback.

Hattie and Timperley (2007) defined feedback as "information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding" (p. 81). In the context of educational assessment, various agents that can provide or receive feedback can be distinguished. For example, a teacher can provide feedback to students, a student can provide feedback to one of his or her peers, or a test result can provide feedback about student learning to a teacher. In early definitions (e.g., Ramaprasad, 1983; Sadler, 1989), feedback required the actual *use* of the information provided. Recent definitions of feedback acknowledge that there are many methods for providing feedback, not all of which are equally effective in terms of contributing to student learning. Moreover, feedback does not necessarily have to come from an external source, but could also come from the learner, which makes it part of the process of self-regulation.

Because the term feedback includes the word "back," it suggests that the process is retroactive. However, in the literature, a distinction is often made between various aspects that constitute good feedback (Hattie & Timperley, 2007): *Feed Up* (where am I going?)*, Feed Back* (how am I going?), and *Feed Forward* (where to next?).

Some forms of feedback include only the *Feed Back* aspect, which indicates a level related to a particular goal or standard, such as a grade or a judgement relating to the correctness of answers to items. However, for feedback to function formatively, it is essential that it provide directions for future learning, such as by indicating strengths and weaknesses, providing explanations, and suggesting next steps to take in the learning process.

1.2.1 Feedback Provided to Students

Feedback provided to students can help reduce the distance between the current and intended learning outcomes (Hattie & Timperley, 2007). Although "there is no best way to provide feedback for all learners and learning outcomes" (Shute, 2008, p. 182), it is possible to identify methods for providing feedback that are generally more effective than others are. However, it must be mentioned that the effects of feedback depend on many factors. Examples are the complexity of the task, the current level and motivational characteristics of the learner. The contents and quality of the feedback can also play an important role (Shute, 2008; Timmers, 2013).

Different responses can occur whenever a student receives feedback. In the ideal situation, the student accepts the feedback and uses it to steer further learning. On the contrary, students might reject or ignore the feedback, which means the feedback will not be used. Another option is that students negotiate feedback, only partly agreeing with the given comments (Stobart, 2008). In order for feedback to be useful and effective for learning, at least three conditions have to be met:

- The learner needs the feedback
- The learner receives the feedback and has time to use it
- The learner is willing and is able to use the feedback

Various categorisations of methods for providing feedback are reported in the literature. However, in order to compare the effects of feedback, a clear classification is needed. In this dissertation, we use the classifications that are currently most widespread in the educational literature and combine the classifications used within two important review studies (Hattie & Timperley, 2007; Shute, 2008). A distinction is made between feedback types (Shute, 2008), feedback levels (Hattie & Timperley, 2007), and feedback timing (Shute, 2008). This classification focuses on item-based feedback and relates specifically to a student's response to an item on a test, instead of the total score achieved on a test, for example.

Feedback types. Shute (2008) describes *knowledge of response* (KR) as a simple type of feedback, which implies that the test taker is told whether the answer is correct or incorrect. This type of feedback is used in behaviourist learning theory, where its main purpose was to reinforce the correct recall of facts (Hattie & Gan, 2011; Narciss, 2008). In the past, KR was often used as a synonym for feedback, indicating that its sole purpose is to inform the learner about the quality of the response (Sadler, 1989). However, the research gradually became aware that a trial-and-error procedure was not very effective in student learning because it does not inform the learner about how to improve. Moreover, in terms of effect sizes, Bangert-Drowns, Kulik, Kulik, and Morgan (1991) concluded that “When learners are only told whether an answer is right or wrong, feedback has virtually no effect on achievements” (p. 228).

A type of feedback that is somewhat more complex is *knowledge of correct response* (KCR), in which the test taker is provided with the correct response (Shute, 2008). This type of feedback originated in cognitivism, the main purpose of which is to revise the student’s incorrect responses (Kulhavy & Stock, 1989).

Any feedback that is more elaborated than KR or KCR is called *elaborated feedback* (EF) (Shute, 2008). In EF, the distinctions between feedback (in terms of correctives) and instruction fades (Hattie & Timperley, 2007), because “the process itself takes on the forms of new instruction, instead of informing the student solely about correctness” (Kulhavy, 1977, p. 212). However, the degree of elaboration strongly differs in various studies, and many variations are possible, not all of which are equally effective. An example of EF is the explanation of the correct answer, a worked-out solution, or a reference to study material. EF plays an important role in recent learning theories (Thurlings, Vermeulen, Bastiaens, & Stijnen, 2013). For example, in social cultural theory, feedback often takes the form of a prompt, which could be classified as EF when students are offered help, such as a hint that guides the learner in the right direction. In (social) constructivism, EF typically includes explanations of strategies or procedures that serve as toolkits in the construction of knowledge and skills (Mory, 2004). In meta-cognitivism, EF is usually concerned with *how* the learner learns, instead of *what* the learner learns (Brown, 1987; Stobart, 2008). These learning theories advise that feedback should be task-related, specific, and objective-oriented (Thurlings et al., 2013).

Feedback levels. Hattie and Timperley (2007) argued that the effectiveness of feedback depends on the level at which the feedback is aimed. They distinguish four levels, which are an expansion of a previously developed model by Kluger and DeNisi (1996):

- *Task level.* The feedback is aimed at correcting work and is mainly focussed on lower-order learning outcomes.
- *Process level.* The feedback relates to the process that was followed in order to finish the task and to how learning can be improved. For example, an explanation is given of why a particular answer is correct.
- *Regulation level.* The feedback is related to the processes in the learner’s mind, such as self-assessment, willingness to receive feedback, self-confidence, and help-seeking behaviour.

- *Self-level.* The feedback is not related to the task but is aimed at the characteristics of the learner. Praise is an example of feedback at the self-level. Feedback at this level is generally thought of as ineffective in learning, although it is the most common level of feedback provided by teachers (Hattie & Timperley, 2007).

Feedback timing. With regard to timing, Shute (2008) distinguished immediate and delayed feedback. Immediate feedback is usually provided immediately after the student answers the item. The definition of "delayed" is more difficult to make, since the degree of delay varies widely among different studies. In some cases, the feedback is delayed until a block of items has been completed. Delayed feedback could also mean that feedback is provided after the student has completed the entire test. However, it is also possible that feedback is provided a day after completion of the test, or even later. Although the results with regard to feedback timing are highly inconsistent, Shute suggested that feedback timing should be based on the intended levels of learning outcomes. Namely, when the feedback is intended to facilitate lower-order learning outcomes, such as the recall of facts, immediate feedback works best. However, when higher-order learning outcomes are at stake and require transfer of what has been learned to a new situation, it is probably best to provide delayed feedback.

An integrated feedback classification model. We connected the theories of Shute (2008) and Hattie and Timperley (2007) in order to derive a comprehensive view of the different methods used to provide feedback (Figure 1.1). The content of the feedback is determined by its type and level. Knowledge of results (KR) and knowledge of correct response (KCR) relate only to the task level, given that they merely provide information concerning the correctness of the student's answer. As indicated above, the nature of elaborated feedback (EF) can vary widely. Therefore, EF can be aimed at all possible levels. Because EF on the self-level is not deemed an effective strategy, the relationship between EF and the self-level is represented by a thin line in Figure 1.1. In addition to the content of the feedback, timing (Shute, 2008) plays an important role. Feedback can be provided either immediately or in a delayed interval after the student has responded to the item.

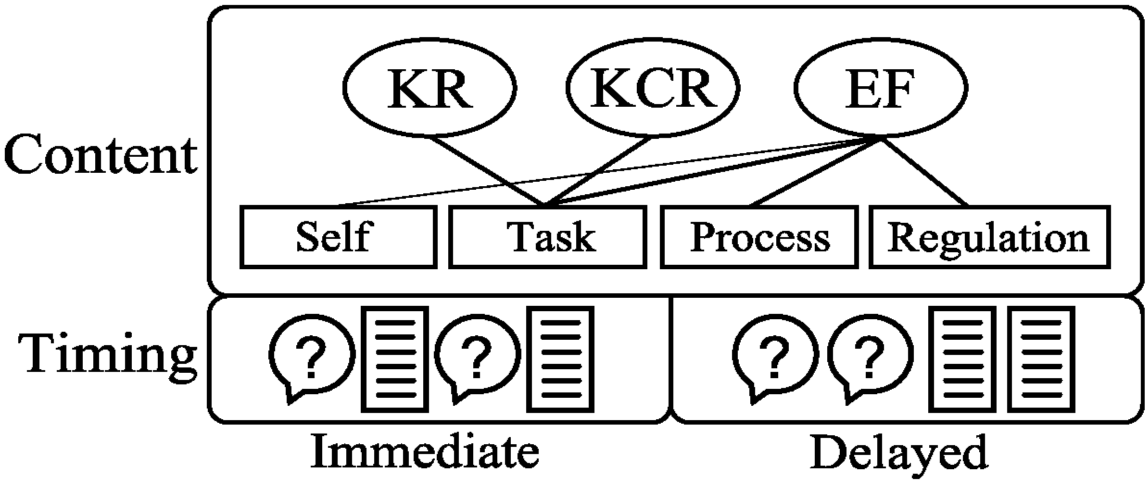


Figure 1.1. Types of feedback (Shute, 2008) linked to levels of feedback (Hattie & Timperley, 2007) and timing (Shute, 2008).

1.2.2 Feedback Provided to Educators

Many tests, such as those given in pupil-monitoring systems, provide users with feedback about pupil performance in the form of a score report at the levels of the individual pupil, class, or school. This is called data feedback. Reports often fulfil a summative function, which means that they provide information regarding the attainment of a certain level or standard. However, when the data feedback is intended to inform learning, it is considered to serve a formative purpose.

Feedback level and content. The feedback in score reports relates to one or multiple test-taking moments, and the level at which the results are reported is usually quite general. This is possibly because most reports have generated feedback regarding student learning in large-scale assessment programs (e.g., National Assessment of Educational Progress in Hambleton & Slater, 1997). For example, overall scores are often reported in terms of percentile rankings, number of correct scores, and IRT-scaled scores (Goodman & Hambleton, 2004). However, some tests provide subscores that allow for a detailed examination of student performance. Assessment reports have frequently been criticised as presenting too much information in a way that is hard for users to interpret (e.g., Goodman & Hambleton, 2004; McMillan, 2001). For a report to be interpretable, it needs to focus on a limited number of purposes, and the information must be displayed clearly and in an uncluttered manner (Goodman & Hambleton, 2004). Various strategies can be used to support the interpretation of the information in the reports, such as providing marks or comparing a student's or group's performance using anchor points, performance standards, market-basket displays, and benchmarks (Jaeger, 2003; Mislevy, 1998). These help the user to understand the test results in terms of both content-referenced and criterion-referenced interpretations (Zenisky & Hambleton, 2012). In addition, providing both graphical representations and numerical information, in either tables or graphs, or a combination of both, can support the interpretation.

Score reports should be carefully designed in consideration of various relevant aspects, such as the proposed uses of the test, the level of the reporting unit (individual student, class, school, state, country, etc.), the desired level of specificity of the report, and the purpose of the particular report. For example, it would be very useful to provide subscores for various aspects of the test to an individual student so that the strengths and weaknesses of that student could be examined. This would improve the formative potential of the report. Nevertheless, providing such subscores in a report that provides information on the performance of an entire class would probably overwhelm the user. Hence, it is often appropriate to offer multiple reports that focus on a particular level and have a specific reporting purpose (e.g., reporting performance related to a certain standard, or reporting growth). It is also possible that overviews of correctly and incorrectly answered items could be provided. Supporting users in interpreting pupils' assessment score reports has recently been addressed as an important aspect of test validity because it is a precondition for the appropriate use of the test results (Hattie, 2009; Ryan, 2006; Zenisky & Hambleton, 2012).

Feedback timing. There is a wide range in possibilities with respect to timing in the provision or availability of reports. Whenever the test has been administered through a computer, it is often possible to generate reports immediately at the individual level. Nevertheless, the time between taking the test and the availability of feedback on the test results is often longer in large-scale testing programs. The length of the feedback loop should be appropriate given the intended uses of the test results. The quick provision of feedback on the results is the most important when providing information about the performances of individual students because their abilities change continuously. When the test results are used only for making decisions at a higher aggregation level, it is generally acceptable that the feedback loop covers a longer time span (Wiliam et al., 2013).

1.3 Computers in Educational Assessment

The interest in the use of computers to support or administer educational assessments has increased rapidly over the last decades. Indeed, using computers in educational assessment can have practical advantages, but more importantly, they offer pedagogical advantages. In this section, the potential of computers for formative assessment are outlined.

1.3.1 Computer-based Assessment

Computer-based assessment (CBA) or computer-based testing (CBT) is a form of assessment where students take a test in a computer environment. Since the term CBT has mainly been used to refer to assessment for summative purposes (e.g., Association of Test Publishers, 2000; The International Test Commission, 2006), the term CBA will be used in this dissertation because it has often been associated with computer-based learning tools and assessments that aim to serve formative purposes.

CBA has practical advantages over paper-and-pencil tests, the most important of which are higher efficiency, reduced costs, higher test security, and the possibility of applying automatic scoring procedures (Lopez, 2009; Parshall, Spray, Kalohn, & Davey, 2002). These advantages are particularly useful in large-scale summative assessments, which makes CBA attractive to use in large educational institutions, since the test can be administered to a large group. Hence, most CBAs have been developed for use in higher education (Peat & Franklin, 2002). However, if they are constructed and implemented adequately, CBAs also have pedagogical advantages (Jodoin, 2003). The pedagogical advantages that apply to both assessments with a summative and formative function are as follows:

- The possibility of quick feedback about student performance (Charman, 1999)
- Flexibility with regard to the location and time the assessment is taken (Charman, 1999; Kearney, Fletcher, & Bartlett, 2002)
- The opportunity to use innovative item types, which allows for more authentic measurements (Parshall et al., 2002; Scalise & Gifford, 2006)
- The possibility of adapting the difficulty of the items in the test to the ability of the students, namely computerized adaptive testing, also known as CAT (Wainer, 2000).

In assessments with a formative purpose, the advantages of CBA are mainly related to the timing and speed of automatically generated (elaborated) feedback (Clements, 1998) and the flexibility of the tests in terms of item selection. The fact that immediate feedback can be provided to students while they are taking the test could lead to higher learning outcomes. As in the one-to-one tutoring situation, which is claimed to be the most effective form of instruction (Bloom, 1984), feedback in CBA can serve to resolve immediately the gap between the student’s current status in the learning process and the intended learning outcome (Hattie & Timperley, 2007). Moreover, this feedback can be provided to each test taker, based on their particular response to an item. This would be unimaginable in a classroom with one teacher and 30 pupils. Furthermore, previous research has suggested that feedback provided to students through computers, as opposed to humans, shows particularly large effects on student learning (e.g., Hattie, 1999; Kluger & DeNisi, 1996). This then may relate to the frequency and timing of feedback; moreover, the feedback provided by the computer may be less threatening than feedback provided by the teacher (Blok, Oostdam, Otter, & Overmaat, 2002).

1.3.2 Computer-based Reporting

If data derived from assessment are used to inform decisions within the school, it is of utmost importance to analyse these data for observing trends or patterns within or across groups of pupils. Assessment data can be very informative in analysing the performance of individual pupils over a certain time span. The analysis of students’ learning outcomes is an important aspect of DBDM, in which high-quality data from standardised tests are often used to support decision-making processes (Schildkamp & Kuiper, 2010). However, paper-based reports of such large-scale assessment programmes have not appeared very helpful for those purposes. Moreover, these data have been typified as “autopsy data,” because of their late availability, lack of instructional relevance, as well as the impossibility of performing additional analyses on the data (Mandinach & Jackson, 2012). Computer-based reporting can possibly help overcome such problems.

Numerous statistical packages are available, which support the analysis of assessment data. However, the degree to which educators are capable of correctly analysing and subsequently interpreting data using such systems is questionable. Numerous studies 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, Blok, Boogaard, & Krüger, 2009; Meijer, Ledoux, & Elshof, 2011; Saunders, 2000; Van Petegem & Vanhoof, 2004; Williams & Coles, 2007; Zupanc, Urank, & Bren, 2009).

Because data interpretation is necessary to alter conditions to meet pupils' needs, it touches upon one of the basic skills that comprise 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 use their intuition to interpret the assessment results. O'Malley, Lai, McClarty, and Way (2013) suggested that technology could help in communicating assessment results because users can demand the information that is relevant to them. Thus, current technology makes it possible to generate reports that are tailored to the needs of particular users. For example, in the USA, data dashboards have become popular tools that automatically create score reports. Regarding interpretation and feedback of test results, the International Test Commission (2006) formulated guidelines specifically aimed at reporting results gathered using computer-based tests. These guidelines state that various reports should be available for different stakeholders. An example of computer-based reporting concerns school performance feedback systems (SPFS), which are systems developed by professional organisations that aim to provide schools with insight into the outcomes of the education they have provided (Visscher & Coe, 2002). Pupil-monitoring systems, a kind of SPFS, have been developed primarily to monitor the individual progress of pupils. These systems allow for the automatic generation of reports at various levels within the school, covering different time spans and test content. These tools reduce the demands placed upon users in terms of statistical skills because they do not have to engage in complex statistical analyses. Nevertheless, there is little knowledge about the degree to which users are capable of correctly interpreting the reported results of assessments, which is a crucial precondition for DDDM.

1.4 Outline

This dissertation covers three areas: 1) item-based feedback provided to students through a computer; 2) feedback provided through a computer to educators based on students' assessment results; 3) comparison of three approaches to formative assessment: data-based decision making (DBDM), assessment for learning (AfL), and diagnostic testing (DT).

1.4.1 Item-based Feedback Provided to Students through a Computer

Chapter 2 presents an experiment that has been conducted at a higher education institute in the Netherlands, focusing on the effects of written feedback in a computer-based assessment of students' learning outcomes. The literature shows conflicting results with regard to the effects of different ways of providing feedback concerning students' learning outcomes (e.g., Kluger & DeNisi, 1996; Shute, 2008). However, regarding written feedback in a CBA, generally, positive effects have been reported for EF aimed at the task and process levels or task and regulation levels. The results with regard to the timing of feedback vary widely (Mory, 2004). Therefore, in the experiment presented in Chapter 2, it was decided to compare the effects of KCR + EF to KR as well as the effects of immediate and delayed feedback. It was expected that students would benefit more from KCR + EF than from KR only, with respect to learning outcomes. Furthermore, it was investigated whether time spent on reading feedback would be roughly the same for identical feedback content, but under different feedback timing conditions. In addition, the relationship between students' attitudes and time spent reading feedback was explored.

Chapter 3 focuses on a systematic review of the effects of written item-based feedback in a computer-based assessment on students' learning outcomes. The purpose of this study was to investigate the effectiveness of different methods for providing feedback in a CBA on students' learning outcomes as well as to identify gaps in current knowledge on this topic. In analysing the results, feedback characteristics such as type, level, and timing were taken into account. The level of learning outcomes was also taken into account, as Smith and Ragan (2005) claimed that different ways of providing feedback are differentially advantageous for certain levels of learning outcomes. The conclusions of the different studies were brought together and synthesized using a qualitative method (narrative analysis).

Chapter 4 presents a meta-analysis aimed at gaining insight into the effectiveness of various methods for providing item-based feedback in a computer-based environment on students' learning outcomes. The currently available evidence in the literature on the effects of feedback on student learning in computer-based environments is limited (e.g., Azevedo & Bernard, 1995; Jaehnig & Miller, 2007; Van der Kleij, Timmers, & Eggen, 2011). Therefore, there is a need for a meta-analysis that takes into account variables that seem relevant given the literature, which builds on some of the conclusions drawn in existing overview studies. Conducting a meta-analysis makes it possible to detect patterns or effects that are not visible at the level of individual experiments. It also provides us with insights into the magnitude of the feedback effects. While in Chapters 2 and 3 only written feedback was considered, in Chapter 4 no restrictions were made with respect to feedback mode.

1.4.2 Feedback Provided through a Computer to Educators

Chapters 5 and 6 focus on how to provide feedback effectively to educators in the form of a computerised score report based on students' assessment results. This research has been conducted in the context of the reports generated by the Computer Program LOVS. This computer program automatically generates reports of test results belonging to Cito's pupil-monitoring system, LOVS.

The purpose of the study presented in Chapter 5 was to (a) investigate the extent to which the reports from the Computer Program LOVS are correctly interpreted by teachers, internal support teachers, and school principals 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, such as training in the use of the Computer Program LOVS and the degree to which the information from the Computer Program LOVS is perceived as useful.

Chapter 6 investigated how the reports generated by the Computer Program LOVS could be redesigned to support users in interpreting pupils' test results. In several rounds of consultations with users and experts, alternative designs for the reports were created and field tested. The aims of this study were twofold. First, solve a problem in practice, i.e., users, particularly teachers, seem to experience difficulties in interpreting the reports generated by the Computer Program LOVS (based on the results of Chapter 5). Second, contribute to the theoretical understanding regarding score reporting.

1.4.3 A Comparison of three Approaches to Formative Assessment

Chapter 7 relates to the broader approach in which computer-based assessments may be used to support learning. The chapter addresses the theoretical differences and similarities amongst three approaches to formative assessment that are currently most frequently discussed in educational research literature. The first approach is *data-based decision making* (DBDM), which defines improving students' learning outcomes in terms of results and attaining specified targets (Wayman et al., 2013). Second, *assessment for learning* (AfL), (Assessment Reform Group [ARG], 1999), is an assessment approach that focuses on the quality of the learning process, rather than merely on students' (final) learning outcomes (Stobart, 2008). Finally, *diagnostic testing* (DT) is an approach in which detailed assessment data about a learner's problem solving are collected to explain his or her learning process and learning outcomes (Crisp, 2012; Keeley & Tobey, 2011). Within some of the approaches, the terminology and definitions are inappropriately used interchangeably; therefore, it is valuable to review and compare the theoretical underpinnings of DBDM, AfL, and DT. To move the field of educational assessment forward, clarity on the theoretical underpinnings is necessary. The study compared the similarities and differences in the theoretical underpinnings of DBDM, AfL, and DT. Furthermore, it explored the consequences of these similarities and differences for implementing DBDM, AfL, and DT in educational practice.

Table 1.1 presents a schematic outline of this dissertation. The chapters in this dissertation can be read independently of one another. To conclude with, an epilogue that is built upon the contents of the individual chapters is presented, and some suggestions for future research are provided.

Table 1.1
Outline of the Chapters in this Dissertation

Chapter	Aspect of formative assessment	Method	Research objectives
2	Written item-based feedback provided to students in a CBA.	Pre-test/post-test design with random assignment, 2 experimental groups and 1 control group.	• Studying the effects of immediate KCR + EF, delayed KCR + EF, and delayed KR on students' learning outcomes in an experimental setting. • Studying the time spent reading feedback.
3	Written item-based feedback provided to students in a CBA.	Systematic review (narrative analysis).	• Investigating the effectiveness of different methods for providing feedback in a CBA. • Identifying gaps in current knowledge on this topic.
4	Item-based feedback provided to students in a computer-based learning environment.	Meta-analysis.	• Gaining insight into the effectiveness of various methods for providing item-based feedback in a computer-based environment on students' learning outcomes.
5	Feedback provided through a computer to educators based on students' assessment results.	Multi-methods design; focus group meetings and consultations with experts (qualitative), and a questionnaire (quantitative).	• Investigating the extent to which the reports from the Computer Program LOVS are correctly interpreted by educators. • To identify various potential stumbling blocks with regard to the interpretation of the reports.
6	Feedback provided through a computer to educators based on students' assessment results.	Educational design research using multiple methods; focus group meetings, consultations with experts, and key informants consultations (qualitative), and a questionnaire and key informants consultations (quantitative).	• Investigating how the reports from the Computer Program LOVS can be redesigned in a way that supports users in interpreting pupils' test results. • Designing alternative score reports for the Computer Program LOVS.
7	Three approaches to formative assessment: data-based decision making (DBDM), assessment for learning (AfL), and diagnostic testing (DT).	Theoretical comparison.	• Investigating the similarities and differences in the theoretical underpinnings of DBDM, AfL, and DT. • Investigating the consequences of these similarities and differences for implementing DBDM, AfL, and DT in educational practice.

References

Assessment Reform Group. (1999). *Assessment for learning: Beyond the black box.* Retrieved from http://assessmentreformgroep.files.wordpress.com/2012/01/beyond_blackbox.pdf

Association of Test Publishers. (2000). *Guidelines for computer-based testing*. Washington, DC: Author.

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research 13*, 111–127. doi:10.2190/9LMD-3U28-3A0G-FTQT

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213–238. doi:10.3102/00346543061002213

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

Black, P., & Wiliam, D. (1998a). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–48.

Black, P., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. London, UK: King's College.

Black, P., & Wiliam, D. (1998c). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*, 7–74. doi:10.1080/0969595980050102

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*, 5–31. doi:10.1007/s11092-008-9068-5

Blok, H., Oostdam, R., Otter, M. E., & Overmaat, M. (2002). Computer-assisted instruction in support of beginning reading instruction: A review. *Review of Educational Research*, *72*, 101–130. doi:10.3102/00346543072001101

Bloom, B. S. (1984). The search for methods of group instruction as effective as one-to-one tutoring. *Educational Leadership*, *41*(8), 4–17. Retrieved from http://www.ascd.org /ASCD/pdf/ journals/ed_lead/el_198405_bloom.pdf

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning.* New York, NY: McGraw-Hill.

Briggs, D. C., Ruiz-Primo, M. A., Furtak, E., Shepard, L., & Yin, Y. (2012). Meta-analytic methodology and inferences about the efficacy of formative assessment. *Educational Measurement: Issues and Practice*, *31*, 13–17. doi:10.1111/j.1745-3992.2012.00251.x/full

Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. E. Weinart & R. H. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 65–116). Hillsdale, NJ: Lawrence Erlbaum.

Brookhart, S. B. (2007). Expanding views about formative classroom assessment: A review of the literature. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 43–62). New York, NY: Teachers College Press.

Brookhart, S. M., & Nitko, A. J. (2008). *Assessment and grading in classrooms.* Upper Saddle River, NJ: Pearson Education.

Charman, D. (1999). Issues and impacts of using computer-based assessments (CBAs) for formative assessment. In S. Brown, J. Bull, & P. Race (Eds.), *Computer-assisted assessment in higher education* (pp. 85–93). London, UK: Kogan Page.

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review, 24*, 205–249. doi:10.1007/s10648-011-9191-6

Clements, D. H. (1998). Computers in mathematics education assessment. In G. W. Bright, & J. N. Joyner, *Classroom assessment mathematics: Views from a national science foundation working conference* (pp. 153–159). Lanham, MD: University press of America. Retrieved from http://investigations.terc.edu/library/bookpapers/computers_in_mathematics.cfm

Crisp, G. T. (2012). Integrative assessment: Reframing assessment practice for current and future learning. *Assessment & Evaluation in Higher Education*, *37*, 33–43. doi:10.1080/02602938.2010.494234

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

Goodman, D., & Hambleton, R. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*, 37–41. doi:10.1207/s15324818ame1702_3

Hambleton, R. K., & Slater, S. C. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Technical Report 430). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Teaching.

Harlen, W. (2007). *The quality of learning: Assessment alternatives for primary education. Interim Reports*. Cambridge, UK: University of Cambridge. Retrieved from http://gtcni.openrepository.com/gtcni/bitstream/2428/29272/2/Primary_Review_Harlen_3-4_briefing_Quality_of_learning_-_Assessment_alternatives_071102.pdf

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. doi:10.2190/ET.36.2.g

Hattie, J., & Gan, M. (2011). Instruction based on feedback. In P. Alexander & R. E. Mayer (Eds.), *Handbook of research on learning and instruction* (pp. 249–271). New York, NY: Routledge

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112. doi:10.3102/003465430298487

Hattie, J. (2009). Visibly learning from reports: The validity of score reports. *Online Educational Research Journal*. Retrieved from http://www.oerj.org/View?action=viewPDF&paper=6

Jaehnig, W., & Miller, M. L. (2007). Feedback types in programmed instruction: A systematic review. *Psychological Record, 57*(2), 219–232.

Jaeger, R. M. (2003). *NAEP validity studies: Reporting the results of the National Assessment of Educational Progress.* Working paper 2003–11. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement, 40*, 1–15. doi:10.1111/j.1745-3984.2003.tb01093.x

Johnson, M., & Burdett, N. (2010). Intention, interpretation and implementation: Some paradoxes of assessment for learning across educational contexts. *Research in Comparative and International Education, 5*, 122–130. doi:10.2304/rcie.2010.5.2.122

Kearney, J., Fletcher, M**.**, & Bartlett, B. (2002). Computer-based assessment: Its use and effects on student learning. In H. Middleton, (Ed.), *Learning in Technology Education: Challenges for the 21st Century* (pp. 235–242*)*. Brisbane: Centre for Technology Education Research, Griffith University. Retrieved from http://www98.griffith.edu.au/dspace/bitstream/handle/10072/1455/19796_1.pdf?sequence=1

Keeley, P., & Tobey, C. R. (2011). *Mathematics formative assessment.* Thousand Oaks, CA: Corwin.

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. Retrieved from http://ld6ela.edublogs.org/files/2008/07/data-article-Kerr-et-al.pdf

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284. doi:10.1037/0033-2909.119.2.254

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47*, 211–232. doi:10.3102/00346543047002211

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*, 279–308. doi:10.1007/BF01320096

Ledoux, G., Blok, H., Boogaard, 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.

Lopez, L. (2009). *Effects of delayed and immediate feedback in the computer-based testing environment* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3358462)

Mandinach, E. B., & Jackson, S. S. (2012). *Transforming teaching and learning through data-driven decision making.* Thousand Oaks, CA: Corwin.

McManus, S. (2008). *Attributes of effective formative assessment*. Washington, DC: Council of Chief State School Officers. Retrieved from http://www.dpi.state.nc.us/docs/accountability/educators/fastattributes04081.pdf

McMillan, J. H. (2001). *Essential assessment concepts for teachers and administrators*. 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, the Netherlands: SCO-Kohnstamm Instituut.

Mislevy, R. J. (1998). Implications of market-basket-reporting for achievement level setting. *Applied Measurement in Education, 11*, 49–63. doi:10.1207/s15324818ame1101_3

Mory, E. H. (2004). Feedback research revisited. In D. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 745–783). Mahwah, NJ: Lawrence Erlbaum Associates.

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merril, J. J. G. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 125–144). Mahwah, NJ: Lawrence Erlbaum Associates.

Nicol, D., & McFarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*, 199–218. doi:10.1080/03075070600572090

O'Malley, K., Lai, E., McClarty, K., & Way, D. (2013). Marrying formative, periodic, and summative assessments: I do. In R. W. Lissitz (Ed.), *Informing the practice of teaching using formative and interim assessment: A systems approach* (pp. 145–164). Charlotte, NC: Information Age.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing.* New York, NY: Springer.

Peat, M., & Franklin, S. (2002). Supporting student learning: The use of computer-based formative assessment modules. *British Journal of Educational Technology, 33*, 515–523. doi:10.1111/1467-8535.00288

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*, 4–13. doi:10.1002/bs.3830280103

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.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119–144. doi:10.1007/BF00117714

Sadler, D. R. (1998). Formative Assessment: Revisiting the territory. *Assessment in Education: Principles, Policy & Practice*, *5*, 77–84. doi:10.1080/0969595980050104

Sanders, P. (2011). Het doel van toetsen [The purpose of testing]. In P. Sanders (Ed.), *Toetsen op school* [Testing at school] (pp. 9–20). Arnhem, the Netherlands: Cito. Retrieved from http://www.cito.nl/~/media/cito_nl/Files/Onderzoek%20en%20wetenschap/ cito_toetsen_op_school.ashx

Saunders, L. (2000). Understanding schools' use of 'value added' data: The psychology and sociology of numbers. *Research Papers in Education, 15*(3), 241–258.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment, 6*(4). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1653/1495

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. doi:10.1016/j.tate.2009.06.007

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39–83). Chicago, IL: Rand McNally.

Shepard, L. A. (2005, October). *Formative assessment: Caveat emptor*. Paper presented at the ETS Invitational Conference, The Future of Assessment: Shaping Teaching and Learning, New York, NY.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153–189. doi:10.3102/0034654307313795

Smith, P. L., & Ragan, T. J. (2005). *Instructional design* (3rd ed.). New York, NY: Wiley.

Stiggins, R. J. (2005). From formative assessment to assessment FOR learning: A path to success in standards-based schools. *Phi Delta Kappan, 87*(4), 324–328.

Swan, G., & Mazur, J. (2011). Examining data driven decision making via formative assessment: A confluence of technology, data interpretation heuristics and curricular policy. *Contemporary Issues in Technology and Teacher Education, 11*(2), 205–222. Retrieved from http://www.editlib.org/p/36021

The International Test Commission. (2006). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing, 6*, 143–171. doi:10.1207/s15327574ijt0602_4

Timmers, C. F. (2013). *Computer-based formative assessment: Variables influencing student feedback behaviour.* (Unpublished doctoral dissertation, University of Twente, Enschede, the Netherlands).

Thurlings, M., Vermeulen, M., Bastiaens, T., & Stijnen, S. (2013). Understanding feedback: A learning theory perspective. *Educational Research Review*, *9*, 1–15. doi:10.1016/j.edurev.2012.11.004

Van der Kleij, F. M., Timmers, C. F., & Eggen, T. J. H. M. (2011). The effectiveness of methods for providing written feedback through a computer-based assessment for learning: A systematic review. *CADMO, 19*, 21–39. doi:10.3280/CAD2011-001004

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*(5)*,* 338–353.

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

Wayman, J. C., Spikes, D. D., & Volonnino, M. (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). doi:10.1007/978-90-481-2660-6

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*, 3–14. doi:10.106/j.stueduc.2011.03.001

Wiliam, D., Kingsbury, G., & Wise, S. (2013). Connecting the dots: Formative, interim, and summative assessment. In R. W. Lissitz (Ed.), *Informing the practice of teaching using formative and interim assessment: A systems approach* (pp. 1–19). Charlotte, NC: Information Age.

Williams, D., & Coles, L. (2007). Teachers' approaches to finding and using research evidence: An information literacy perspective. *Educational Research, 49*, 185–206. doi:10.1080/00131880701369719

Wainer, H. (Ed.) (2000). *Computerized adaptive testing. A primer* (2nd ed.). Hilsdale, NJ: Lawrence Erlbaum.

Young, V. M., & Kim, D. H. (2010). Using assessments for instructional improvement: A literature review. *Educational Policy Analysis Archives, 18*(19). Retrieved from http://epaa.asu.edu/ojs/article/view/809

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. doi: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. doi:10.1080/09243450802696695

Chapter 2. Effects of Feedback in a Computer-Based Assessment for Learning[1]

Abstract

The effects of written feedback in a computer-based assessment for learning on students' learning outcomes were investigated in an experiment at a higher education institute in the Netherlands. Students were randomly assigned to three groups, and were subjected to an assessment for learning with different kinds of feedback. These are immediate knowledge of correct response (KCR) + elaborated feedback (EF), delayed KCR + EF, and delayed knowledge of results (KR). A summative assessment was used as a post-test. No significant effect was found for the feedback condition on student achievement on the post-test. Results suggest that students paid more attention to immediate than to delayed feedback. Furthermore, the time spent reading feedback was positively related to students' attitude and motivation. Students perceived immediate KCR + EF to be more useful for learning than KR. Students also had a more positive attitude towards feedback in a CBA when they received KCR + EF rather than KR only.

[1] This chapter is a minor revision of Van der Kleij, F. M., Eggen, T. J. H. M., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers & Education, 58*, 263–272. doi:10.1016/j.compedu.2011.07.020

2.1. Introduction

With the introduction of technology in the classroom, educators have been given a larger number of technological tools to enhance student learning. One of these innovations is computer-based assessment (CBA), a form of assessment in which students answer items in a computer environment instead of taking a traditional paper-and-pencil test. The literature suggests that CBA can have pedagogical advantages because it is possible to provide students with feedback while they are taking the test. This implies that assessment should be integrated into the learning process, which is an important aspect of the *assessment for learning* approach (for more information, see Stobart, 2008). When it comes to assessment for learning, feedback "is seen as the key to moving learning forward" (Stobart, 2008, p. 144).

The fact that feedback can be provided to students in a timely fashion—while they are taking the test—might lead to better learning outcomes. This is because in a computer-based environment, the discrepancies between students' current state and the intended learning outcomes can immediately be solved (Hattie & Timperley, 2007), in contrast to a traditional environment. A big advantage of CBA is the possibility of providing the test taker with customised feedback, given that the computer can generate feedback based on the answer given by the student (Lopez, 2009). This feedback may simply indicate the correct answer for an item or may be more elaborate and provide information concerning the content to which the item refers. Currently available research does not provide univocal evidence regarding how to integrate feedback into a computer-based assessment in such a way that contributes positively to the learning process and to the learning outcomes of students. This study investigated the effects, on students' learning outcomes, of different methods for providing written feedback in a computer-based assessment for learning. Additionally, it explored the attitudes of students towards different methods of providing feedback as well as students' feedback-reading behaviour in terms of time spent reading feedback.

2.1.1 Computer-based Assessment for Learning

Assessment for learning is an approach to classroom assessment in which it is integrated into the learning process (Stobart, 2008). The main aim of assessment for learning is to support the learning process. This is in contrary to the conception that assessments should be used only for summative purposes, a notion some claim leads to teaching to the test (Birenbaum et al., 2006). The assessment for learning approach includes more than a way to use assessments and their results, for example involving students actively in their own learning, adapting teaching in response to assessment results, conducing self and peer assessments and providing students with feedback (for more information, see Assessment Reform Group, 1999; Stobart, 2008). In this research, the focus is on feedback provided to individual students taking part in a computer-based assessment for learning. Feedback is a crucial aspect of assessment for learning because it helps students to gain insight into their present position in the learning process and provides them with information on how to get from their current position to their desired position (Stobart, 2008). In other words, by receiving feedback, the student can adapt his or her learning in order to achieve the desired learning outcomes. However, various studies (Hattie & Timperley, 2007; Shute, 2008, Stobart, 2008) show that feedback does not always contribute positively to the learning process, which emphasises the need for further research on this topic.

With respect to the effects of feedback in a CBA for learning, the results from the literature are mixed. Various authors have reported positive effects on students' learning outcomes as a result of certain methods of providing feedback (Corbalan, Paas, & Cuypers, 2010; Lee, Lim, & Grabowski, 2010; Smits, Boon, & Sluijsmans, 2008; Wang, 2011). In other studies, no effects were found (Clariana & Lee, 2001; Corbalan, Kester, & van Merriënboer, 2009; Gordijn & Nijhof, 2002; Kopp, Stark, & Fischer, 2008). The results of these studies indicate that the characteristics of the feedback intervention and the intended level of learning outcomes are relevant aspects that must be taken into account when examining the effects of CBA feedback on students' learning outcomes. Besides, other variables, such as student's attitudes and motivation, play important roles.

2.1.2 Characteristics of Feedback

Based on a literature study, different types of written feedback in a CBA were distinguished. In her review study, Shute (2008) suggests making a distinction between feedback *type* and feedback *timing*. With regard to feedback types, she makes a distinction based on the *specificity* and *complexity and length* of the feedback. Shute describes knowledge of results (KR) as a relatively low-complexity type of feedback; it only states whether the answer is correct or incorrect. A type of feedback with a higher complexity is knowledge of correct response (KCR); this means the correct response is given whenever the answer is incorrect. A much more complex form of feedback is *elaborated feedback* (EF); however, the degree of elaboration in various studies strongly differs (Shute, 2008). Examples of EF are an explanation of the correct answer, a worked-out solution or a reference to study material.

Furthermore, the timing of feedback plays an important role. Shute (2008) distinguishes immediate and delayed feedback. Immediate feedback is (usually) provided immediately after answering each item. The definition of delayed is more difficult to make, since the degree of delay can vary. In some cases, the feedback is delayed until a block of items has been completed. Delayed feedback could also mean feedback is provided after the student has completed the entire assessment. However, feedback can also be provided an entire day after completion of the assessment or even later. Mory (2004) points out that the claims made with regard to the effects of immediate and delayed feedback vary widely. This variation is, however, strengthened by the fact that the definitions for immediate and delayed feedback vary widely. In the current study, *immediate feedback* is defined as feedback given immediately after completion of an item and *delayed feedback* is defined as feedback given directly after completion of all the items in the assessment.

Hattie and Timperley (2007) distinguish four *levels* at which feedback can be aimed, which is an expansion of a previously developed model by Kluger and DeNisi (1996). The levels distinguished are the self, task, process, and regulation levels. Feedback at the self level is not related to the task performed but is aimed at characteristics of the learner. Praise is an example of feedback at the self level. Feedback at the task level is mainly intended to correct work and is focussed at a surface level of learning (e.g., knowledge or recognition); for example, the student is told whether the answer is correct or incorrect. Feedback at the process level relates to the process that was followed in order to finish the task. In this case, for example, a worked-out example is given. Feedback at the regulation level is related to processes in the mind of the learner, like self-assessment and willingness to receive feedback. In the ideal situation, the feedback is adapted to the current level of the learner (Hattie & Gan, 2011). Hattie and Timperley favour feedback that is aimed at the process or regulation level in order to enhance learning. Feedback at the self level is not seen as effective for learning because it does not provide the student with information regarding how to achieve the intended learning goals.

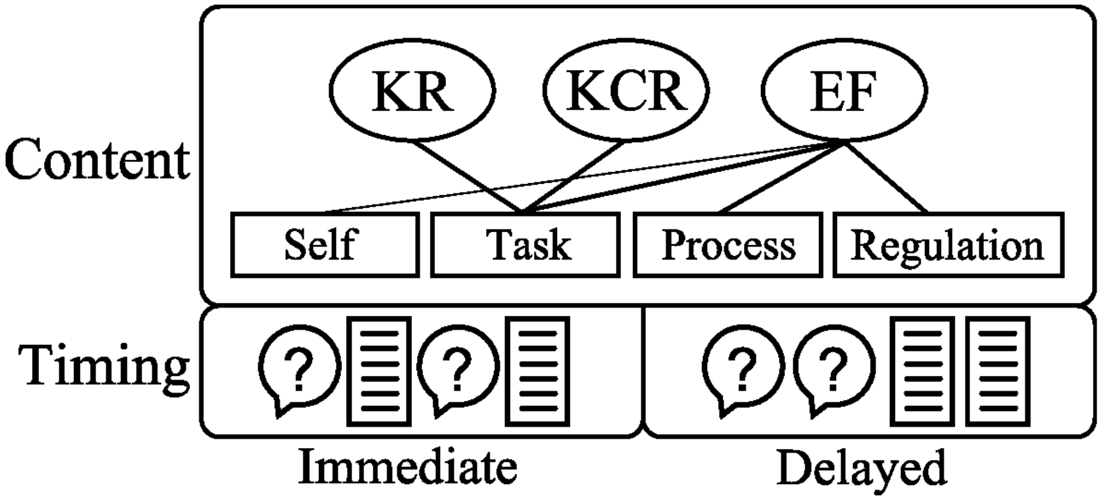


Figure 2.1. Types of feedback (Shute, 2008) linked to levels of feedback (Hattie & Timperley, 2007) and timing (Shute, 2008).

In order to develop a more comprehensive view concerning the different ways of providing feedback, we made a connection between the theory of Shute (2008) and the theory of Hattie and Timperley (2007) (Figure 2.1). The type and level of feedback together determine the content of the feedback. KR and KCR only relate to the task level, given that they merely provide information concerning the correctness of the given answer. As indicated before, the nature of EF can vary widely. Therefore, EF can be aimed at all possible levels. Because EF on the self level is not seen as an effective strategy, the relationship between EF and self level is represented by a thin line in Figure 2.1. Besides the content of the feedback, timing (Shute, 2008) plays an important role. Feedback can be provided either immediately or with delay.

2.1.3 Feedback and Learning

Learning outcomes are the outcomes of the learning process in which the student executed particular tasks (Smith & Ragan, 2005). Even though the effects of various feedback interventions on learning have been investigated to a large extent (e.g., Hattie & Timperley, 2007; Shute, 2008), there is no univocal evidence available due to conflicting results (Kluger & DeNisi, 1996; Shute, 2008). As shown in Figure 2.1, feedback interventions can be classified by type, level, and timing.

The characteristics of the feedback intervention determine to a great extent the effectiveness of the feedback. For example, Clariana, Wagner, and Murphy (2000) reported higher retention and recognition levels for students who received delayed KCR than for students who received immediate KCR or immediate KR with the option to try to solve the item again. This outcome suggests that both the type and timing of the feedback influence the degree to which it affects learning. Lee et al. (2010) compared the effects of immediate EF in the form of metacognitive feedback (at the task and regulation levels) with the effects of immediate KR. They found that EF was more effective than KR both for comprehension and recall tasks, which suggests that the type and level of feedback determine if the feedback is effective.

Besides the characteristics of the feedback, one needs to take various other variables into account that influence the relationship between feedback and learning. Stobart (2008) states that three conditions have to be met in order for feedback to be effective and useful: 1) The learner needs the feedback, 2) the learner receives the feedback and has time to use it and 3) the learner is willing and able to use the feedback. Regarding the first, students need feedback if there is a gap between the current understanding and the goal (Hattie & Timperley, 2007). This implies that if there is no gap, students feel no need to receive feedback. Moreover, Timmers and Veldkamp (2011) showed that it cannot be assumed that all students pay equal attention to feedback provided in a CBA for learning. One aspect that influences attention paid to feedback is the correctness of the answer, with more attention paid to feedback for incorrectly answered items (Timmers & Veldkamp, 2011). Furthermore, their results suggest that with an increase in test length, the willingness to pay attention to feedback decreases. These findings are in line with Stobart's (2008) claim and suggest that the interaction between the difficulty of the item, length of the assessment, and characteristics of the learner determine the amount of attention paid to the feedback, and subsequently the feedback's effect. But, as Stobart points out, the willingness and ability of the student to actually *use* the feedback also plays an important role. The willingness to use feedback is related to students' motivation, which many authors have recognised as an important variable in relation to feedback (see Azevedo & Bernard, 1995; Keller, 1983; Mory, 2004). Additionally, students should be provided enough resources to improve their learning. If, for example, the feedback refers to a source not available to students, they will not be able to use the feedback (Stobart, 2008). Also, the feedback should be presented clearly, with as little distracting information as possible, in order to make it possible for students to successfully process the feedback (Mory, 2004), and subsequently make use of it.

2.1.4 Aims of the Present Study

The literature shows conflicting results with regard to the effects of different ways of providing feedback concerning students' learning outcomes (e.g., Kluger & DeNisi, 1996; Shute, 2008). However, regarding written feedback in a CBA, generally, positive effects have been reported for EF aimed at the task and process levels or task and regulation levels. The results with regard to the timing of feedback vary widely (Mory, 2004). Therefore, in the present study, it was decided to compare the effects of EF to KR as well as the effects of immediate and delayed feedback.

In the study, the effects of immediate KCR + EF, delayed KCR + EF, and delayed KR on students' learning outcomes are investigated. We did not expect a positive effect of KR on learning, given that it does not provide the student with any information on how to improve the current performance. Also, various studies have already shown that students do not benefit from KR (Clariana, et al., 2000; Clariana & Lee, 2001; Kopp et al., 2008; Morrison, Ross, Gopalakrishnan, & Casey, 1995). In terms of effect sizes, Bangert-Drowns, Kulik, Kulik, and Morgan (1991) concluded that "When learners are only told whether an answer is right or wrong, feedback has virtually no effect on achievements (*ES* = -0.08)" (p. 228). For this reason, the control group in this experiment received KR only. The feedback was presented on the same screen as both the item and the students' response, as was advised by Mory (1994) (see Appendix 2A). It was expected that students would benefit more from KCR + EF than from KR only, with respect to learning outcomes (Hypothesis 1).

The contents of the immediate and delayed KCR, both with EF, were identical; only the timing differed. Therefore, it was expected that students in these two feedback conditions would spend roughly the same amount of time reading the feedback but that students who received KR would spend less time reading the feedback because of the shorter feedback length and complexity (Hypothesis 2).

Furthermore, Timmers and Veldkamp (2011) showed that the time spent reading feedback is influenced by different aspects, such as the characteristics of the learner. It was expected that students with a more positive attitude and higher study motivation would spend more time reading feedback (Hypothesis 3).

Hypothesis 1: *Students who receive KCR + EF score significantly higher than students who only receive KR on the summative assessment when controlled for the influence of class and score on the assessment for learning.*

Hypothesis 2: *There is no difference between time spent reading feedback between students who receive immediate or delayed KCR + EF, but students who only receive KR spend less time reading feedback.*

Hypothesis 3: *Student characteristics (attitude, motivation) are positively related to time spent on feedback.*

2.2 Method

2.2.1 Participants

A group of 152 first-year students ($N = 152$, 28 women, 124 men, $M_{\text{age}} = 20.30$ years; $SD = 2.45$, age range: 17 to 27 years) in Commercial Economics (CE) at an institute for higher education participated in this study. The students were selected from nine classes. These students were subjected to an assessment for learning that preceded a summative assessment, the latter of which is part of a Marketing course. The sample selected for this experiment represents 89% of all students who signed up for this period's course exam.

The summative assessment was used as a substitute for the regular paper-and-pencil test and counted for half of the final mark. Additionally, students had to complete a paper-and-pencil test at another point in time, which included a case assignment.

2.2.2 Design

A pre-test/post-test experimental design was used to compare the effects of different ways of providing feedback. The pre-test consisted of an assessment for learning, including different types of feedback for the three groups. Students from the nine classes were randomly assigned to one of the three experimental groups, in which the feedback conditions for the assessment for learning differed. Within each class, students would experience different conditions. The first group (immediate KCR + EF) contained 52 students (14 women, 38 men). The second group (delayed KCR + EF) was composed of 48 students (seven women, 41 men). The third group (delayed KR) contained 52 students (seven women, 45 men). In the next sections, the different types of feedback in the assessment for learning are explained in detail. The post-test consisted of a summative assessment.

The dependent variable in this study was students' scores on the summative assessment. The score on the assessment for learning was used as a measure of previous knowledge; this variable was used as a covariate in order to control for initial differences between students. All students involved in the experiment attended comparable lectures and had access to the same study materials. It was assumed that after random assignment to the experimental conditions, no differences would exist between the three groups. Therefore, possible differences between the groups' scores on the summative CBA could be explained by the different feedback conditions.

2.2.3 Instruments

The instruments used in this experiment were

- an assessment for learning;
- a summative assessment;
- a questionnaire; and
- a time log.

The software used for the administration of the assessment and questionnaire was Question Mark Perception [QMP] (Question Mark Perception, Version 4.0). A screenshot of part of the assessment for learning is presented in Appendix 2A, and a screenshot of the summative assessment can be found in Appendix 2B. Both assessments in the experiment consisted of 30 multiple-choice items with four response options each, which is the regular type of assessment for the course in which the assessment was administered. Moreover, multiple-choice items are easy to score in a CBA and also present practical advantages when analysing the test results. All items and the feedback in the assessment for learning were constructed by teachers of the course from the higher-education institute.[2]

The assessments and questionnaire were administered in the Dutch language because the educational programme of the students was in Dutch. In the subsequent sections, the instruments used in the experiment are elaborated upon.

Assessment for learning. The assessment for learning was intended to support student learning. For this experiment, three different feedback conditions were constructed within the computer-based assessment for learning (see Table 2.1). The contents of the items were identical; however, the type, level, and timing of feedback differed.

The first experimental group was offered a computer-based assessment for learning in which immediate KCR + EF was provided after answering each item. The EF in this experiment gave a concise explanation on how to obtain to the correct answer. Depending on the content of the item, the feedback was aimed at the task or process level according to the classification by Hattie and Timperley (2007). The reason for choosing concise feedback was that Mory's (2004) extensive study showed that simple but sufficient feedback may lead to higher effectiveness with regard to learning than would elaborate feedback. This is because it might contain a considerably smaller amount of distracting information, which makes it relatively easier for students to process the feedback.

The second experimental group was offered the same feedback in the assessment for learning. However, the timing of the feedback differed—they received the feedback after they had completed all the items in the assessment for learning. The other conditions were identical to those of the first experimental group.

The third experimental group served as a control group; students in this group only received feedback on the correctness or incorrectness of the provided answers (KR) after completing the entire assessment for learning. The conditions within the three experimental groups are summarised in Table 2.1.

[2] The experiment was performed at a university of applied sciences in the Netherlands.

Table 2.1
Feedback Conditions Within the Experiment

Group 1	Group 2	Group 3
Immediate KCR + EF	Delayed KCR + EF	Delayed KR
Immediate computer-based written feedback: KCR.	Delayed (after completing the entire assessment for learning) computer-based written feedback: KCR.	Delayed (after completing the entire assessment for learning) computer-based written feedback: KR.
Additional feedback gives an explanation on the correct answer, regardless of whether the answer is correct or incorrect.	Additional feedback gives an explanation on the correct answer, regardless of whether the answer is correct or incorrect.	No additional feedback, the correct answer is not provided.

In constructing the EF, teachers of the subject matter were given guidelines. These guidelines stated that the correct answer had to be provided, accompanied by EF that would give a concise explanation of how to obtain the correct answer. The EF could either be a verbal explanation or a worked-out solution, depending on the nature of the item. The researchers checked if the feedback fulfilled the requirements. The items included different types of tasks, namely knowledge, comprehension, and application. Examples of the items and the accompanying feedback can be found in Appendix 2C.

Summative assessment. The summative assessment was intended to measure student knowledge and understanding of the subject matter. Just like the assessment for learning, the items included different types of tasks, namely knowledge, comprehension, and application.

Questionnaire. The questionnaire is based on one designed by Miller (2009). Her questionnaire was intended to measure students' perceived usefulness of formative CBAs and the extent to which students accept CBAs as a tool for learning. She reported a high reliability of this questionnaire ($\alpha = .92$). Within the current experiment, the questionnaire was mainly intended to measure student motivation, perceived test difficulty, perceived usefulness of the feedback and whether students read the feedback. The questionnaire consisted of items measured on a five-point Likert scale (varying from 1 = *strongly disagree* to 5 = *strongly agree*). The following are examples of items: "I am motivated to learn for this subject", "The difficulty level of the first assessment is too low", "The level of the first assessment is too high", "In general, the feedback in the first assessment was useful", "The feedback was sufficiently elaborate" and "For the items I answered incorrectly, I examined the feedback". At the end of the questionnaire, an open-ended item was added so that students could give comments about the assessment for learning or the summative assessment.

Time log. The amount of time (in seconds) a certain feedback screen was open, was used as an indication of attention paid to feedback. Time logs were also used to investigate whether there was a difference between the behaviours of the three groups. Unfortunately, QMP provided the time a screen with a certain item was open, but the time a certain feedback screen was open was not provided. These data were obtained by subtracting the time spent on completing the items from the total time spent on the assessment for learning. Besides, the questionnaire contained questions about whether students read the feedback.

Checking functionality of the instruments. A pilot test was performed with a small group of students (*N* = 8) enrolled in exactly the same study programme as the participants of the experiment but at a different location. The aim of the pilot test was to investigate if the instruments functioned as intended. Students were asked to provide feedback on problems or mistakes in the assessments.

Additionally, all parts of the assessment were evaluated several times before the assessment was administered at different locations. Some adaptations were made in the assessments and questionnaire after performing the pilot tests; for example, the instruction on the screen was adapted. Furthermore, students reported they did not like the fact that it was not possible to navigate through the items in the assessment for learning with immediate feedback. This implied that students had to start with item one, then move on to item two, etc. We were aware of this disadvantage; however, due to software limitations, we could not change this. In order to keep the conditions within the three groups as identical as possible, it was also decided to not let the other groups navigate in the assessment for learning.

Quality of the assessments. The quality of the assessments was investigated applying Classical Test Theory (CTT) and Item Response Theory (IRT). The software packages TiaPlus (TiaPlus, 2009) and the One Parameter Logistic Model (OPLM) (Verhelst, Glas, & Verstralen, 1995) were used for analysing the data from both a CTT and an IRT perspective.

The assessment for learning was judged to be of sufficient quality based on the quality indicators provided by TiaPlus. For this assessment, Cronbach's alpha α = .85. The summative assessment was judged to be of insufficient quality. For this assessment, α = .40. This value is considered too low, which means the assessment was not reliable. In order to measure the underlying constructs of the assessments, a factor analysis was performed. The result showed that the assessment for learning measured one factor but that the summative assessment measured more than one factor. This meant the summative assessment measured constructs other than the assessment for learning. Therefore, the summative assessment was not a suitable instrument for measuring the learning gains of students within the different groups. In order to overcome this problem, a selection of items was made for the summative assessment based on the factor analysis. These items measured the same construct as the items in the assessment for learning. The number of remaining items was 11. These 11 items together had a reliability of α = .66, which means that removing the other items led to an increase in the reliability of the summative assessment. However, the reliability was still low. From the assessment for learning, one item was removed because it was too easy. Using IRT, the ability of the students (θ) was estimated ($R1c$ = 98.242; df = 78).

There appeared to be no difference between the initial ability of the students in the three groups. Besides the differences in reliability between the two assessments, the results of the CTT and IRT analyses also showed that the summative assessment was more difficult than the assessment for learning.

2.2.4 Procedure

The assessment for learning and the summative assessment were administered immediately after each other; otherwise, the scores on the post-test could be influenced by some intervention other than the feedback. Interaction between students from different groups was not possible, since all students took the assessments at the same time in a supervised environment. While taking the assessments, students were allowed to make notes or calculations on a piece of paper that was provided by the supervisor.

Three weeks before taking the assessments, the teachers informed the students of what they could expect from the assessment session. Additionally, students were sent information about the assessment procedures by e-mail. Also, the teachers kindly requested that the students fill in the questionnaire, which was administered directly after the summative assessment.

On the day the assessments were administered, students received an e-mail with a personal QMP log-in account and a password for the CBA. Students were given two-and-a-half hours to complete the CBAs and the questionnaire. They had to stay in the computer room for at least 45 minutes. These restrictions were put in place to make sure all students would be seriously engaged in the CBA.

2.2.5 Data Analyses

The effects of the feedback in the different conditions were calculated using two-way ANCOVA, which accounted for the sampling of students from classes (Hypothesis 1). The proportion correct on the assessment for learning was used as a covariate in order to control for initial differences between students. The dependent variable in the analysis was the proportion correct on the summative assessment with 11 items ($\alpha = .66$) and was the proportion correct on the assessment for learning with 29 items ($\alpha = .87$).

The total time (in seconds) spent on reading feedback was logged for each student. ANOVA was used to investigate if there was a difference between the three groups' mean times spent on feedback (Hypothesis 2). Furthermore, the results of the questionnaire provided a self-reported measure of time spent reading feedback. This information was used in addition to the time logs in order to investigate if there were differences between the feedback-reading behaviours of students within the three groups.

The questionnaires provided information on relevant student characteristics, such as motivation and attitude towards the CBAs and feedback. A correlation analysis was used in order to investigate the relationship between student characteristics and time spent reading feedback (Hypothesis 3).

In order to measure the underlying constructs of the questionnaire, a factor analysis was performed. The reliability (α) was calculated for each factor measured by the questionnaire. A one-way ANOVA was used to investigate if there were differences between the sum scores of the students within the three groups regarding the factors measured by the questionnaire. Furthermore, post-hoc analyses using the Bonferroni method were performed. Also, a qualitative analysis was performed on the results of the questionnaires. Students' responses to individual items in the questionnaire were analysed using bar charts and cross tabulations. The dispersion of response patterns was analysed and reported.

2.3 Results

2.3.1 Feedback Effects

Two students did not complete both assessments; therefore, data from these students were not taken into consideration in analysing the feedback effects. The remaining students' scores (expressed in proportion correct) on the assessment for learning and the summative assessment were explored and compared. Proportions correct ranged from .24 to .97 in the assessment for learning and from .18 to 1.00 in the summative assessment. Table 2.2 shows the results of the comparisons of the mean proportions correct for the three groups.

Table 2.2
Proportions Correct for the three Groups and Time Spent Reading Feedback

Group	Participants	Proportion correct Assessment for learning		Proportion correct Summative assessment		Time log (seconds)	
	n	*M*	*SD*	*M*	*SD*	*M*	*SD*
Group 1	52	.62	.20	.63	.22	257.60	129.38
Group 2	47	.65	.19	.66	.21	138.60	149.09
Group 3	51	.68	.21	.65	.22	96.61	75.30
Total	150	.65	.20	.65	.22	165.57	138.95

From Table 2.2, it can be concluded that students in all groups scored comparably on the assessment for learning and the summative assessment. Also, the standard deviations were large for all groups. Levene's test for equality of variances shows that the groups are homogenous: $F(2, 147) = 0.25, p = .775$.

At first, students' proportions correct on the summative assessment were compared among the three groups using one-way ANCOVA. The proportion correct on the assessment for learning was added as a covariate in order to control for previous achievement. The one-way ANCOVA, $F(2, 149) = 13.99$, $p = .822$, $\eta^2 = 0.003$, demonstrated no significant differences between the groups regarding the proportions correct on the summative assessment.

In order to investigate whether the proportions correct on the summative assessment differed between the *classes*, proportions correct of classes were compared using one-way ANOVA. For both the assessment for learning, $F(8, 149) = 12.92$, $p < .001$, $\eta^2 = .42$, and summative assessment, $F(8, 149) = 13.99$, $p < .001$, $\eta^2 = .44$, it was shown that there were differences between the classes.

This indicates that within this experiment, there was a difference between the class means, and this should be accounted for in the analyses. Subsequently, it was investigated whether the differences on the summative assessment were still present after correcting for the proportions correct on the summative assessment. The one way ANCOVA, $F(8, 149) = 6.42$, $p < .001$, $\eta^2 = .27$, demonstrated significant differences *between* the proportions correct on the summative assessment of the students in the nine classes, controlling for achievement on the assessment for learning. This indicates that some of the differences between classes cannot be explained by differences on the pre-test.

Hypothesis 1: *Students who receive KCR + EF score significantly higher than students who only receive KR on the summative assessment when controlled for the influence of class and score on the assessment for learning.*

Next, Hypothesis 1 was tested. The aim of this study was not to investigate class differences but to investigate the effects of different ways of providing feedback. Thus, we were not so much interested in the class differences but in the group differences. The fact that the mean proportions correct on both assessments differed between classes is therefore a disturbing variable. In order to take the class differences into account, a two-way ANCOVA was performed. Here, the effects of the groups on the proportion correct on the summative assessment were investigated, taking into account that the means for the classes differed. Also, the proportions correct on the assessment for learning were included as a covariate. Since ANCOVA assumes linearity of the regression lines, first it was investigated, using ANOVA, if there was an interaction effect of groups and classes. The ANOVA showed that there was no interaction effect (p = .80), which means the assumption of linearity was not violated. The two-way ANCOVA showed that when the differences between classes were accounted for, the feedback condition did not significantly affect students' achievement on the summative assessment, $F(2, 138) = 0.11$, $p = .89$. Therefore, Hypothesis 1 was rejected.

Even though no effects of feedback were found on learning, the questionnaire provided valuable information on students' opinions with regard to the different feedback conditions. Namely, the qualitative analysis of the questionnaire showed that that the opinion of students concerning the usefulness of CBAs for learning differed between the three groups. Students who received immediate or delayed KCR + EF were more positive than those who received delayed KR. Also, students who received immediate KCR + EF were more positive than those who received delayed KCR + EF. A comparable pattern in the opinion of students was observed regarding the degree to which students indicated that they learned from feedback in a CBA: again, students who received immediate KCR + EF were more positive than those who received delayed KR. No differences were present between students who received delayed KCR + EF and delayed KR. These results suggest that students prefer immediate feedback to delayed feedback. Furthermore, large differences were present among students' responses with regard to the usefulness of the feedback. Students who received KCR + EF agreed that that the feedback was useful, while the opinions of students who received KR were more diverse and more negative.

2.3.2 Time Spent Reading Feedback

It was expected that students who received KCR + EF (Groups 1 and 2) would spend about the same amount of time reading the feedback in the assessment for learning, given that the contents of the feedback are identical. The feedback in Group 3 showed only KR, which does not take a lot of time to examine because of its short length and low complexity. Therefore, it was expected that students in Group 3 would spend less time reading the feedback than would students in Groups 1 and 2.

Hypothesis 2: *There is no difference between time spent reading feedback between students who receive immediate or delayed KCR + EF, but students who only receive KR spend less time reading feedback.*

For each student, the total time (in seconds) spent reading the feedback was calculated. The means for the three groups can be found in Table 2.2.

Table 2.2 shows that students in Group 1 spent the most time reading the feedback, followed by Group 2 and then Group 3. In order to investigate if there was a significant difference between the groups regarding time spent reading feedback, an ANOVA was performed. The results show that not all group means were the same: $F(2, 147) = 24.40$, $p < .001$, $\eta^2 = .25$. Post-hoc analysis shows that the mean time spent reading feedback differed significantly between Group 1 compared to Groups 2 and 3 ($p < .001$). The difference between the means of Groups 2 and 3 was not significant ($p = .266$). Based on the outcomes of the ANOVA, Hypothesis 2 was rejected.

Additionally, the questionnaire provided information regarding students' feedback-reading behaviour. Students indicated that they paid more attention to immediate feedback than to delayed feedback. Students who received immediate KCR + EF were more likely to read the feedback whenever they guessed an item than were students who received delayed KCR + EF or KR only. Also, results suggest that students paid more attention to feedback for incorrectly answered items than for correctly answered items. The results from the qualitative analysis of the questionnaire supported the outcomes of the analysis of the time logs.

2.3.3 Student Characteristics and Perceived Test Difficulty

The results of the factor analysis showed that the questionnaire measured two factors. Factor 1 included 11 items ($\alpha = .84$). Factor 2 included four items ($\alpha = .78$). Factor 1 included items that measured student characteristics, namely their attitude towards the assessments and feedback in CBAs. Factor 2 measured students' perceived difficulty of the assessments. Using PP-plots, it was investigated whether the responses were normally distributed. Small deviations from normal were found, but no serious deviations were discovered.

The differences between the sum scores of the groups on Factor 1 and 2 were analysed using ANOVA. The results show that the difference for Factor 1 is significant, $F(2, 148) = 7.45$, $p = .001$, $\eta^2 = .28$. Post-hoc analysis shows that the factor scores differ significantly between Groups 1 and 3 ($p = .001$) and between Groups 2 and 3 ($p = .035$). The difference between Groups 1 and 2 is not significant ($p = .743$). These outcomes suggest that students have a more positive attitude towards feedback in a CBA when they receive KCR + EF rather than KR only.

No significant differences were found between the three groups for Factor 2, $F(2, 148) = 0.49$, $p = .613$. This means that there were no differences between the three groups regarding the perceived difficulty of the assessments. Student motivation was about equal for all groups—the majority of the students agreed that they were motivated to learn the subject.

2.3.4 Relation between Student Characteristics and Time Spent on Feedback

It was expected that students with a more positive attitude and greater study motivation would spend more time reading feedback (Hypothesis 3).

Hypothesis 3: *Student characteristics (attitude, motivation) are positively related to time spent on feedback.*

A two-tailed Pearson correlation analysis was performed in order to investigate the relationship between students' attitudes towards CBAs for learning and time spent reading feedback. The relationship was found to be moderately positive and significant at $\alpha = .01$, $r(150) = .32$, $p < .01$. Also, a correlation analysis was performed concerning study motivation and time spent reading feedback. The relationship was slightly positive but significant, $r(150) = .20$, $p < .05$. These outcomes show that both students' attitudes and motivation were related to the time spent reading feedback, which implies Hypothesis 3 was not rejected.

2.4 Discussion

In this study, an experiment was conducted to investigate the effects of different types of written feedback in a computer-based assessment for learning on students' learning outcomes. Students were randomly assigned to one of three experimental groups and were all subjected to an assessment for learning, summative assessment, and questionnaire. The contents of the assessments were identical for all groups, except for the feedback in the assessment for learning. The effects of immediate KCR + EF (Group 1), delayed KCR + EF (Group 2), and delayed KR (Group 3) were compared.

We had hypothesised that students in Groups 1 and 2 would score significantly higher on the summative assessment than would students in Group 3 when controlled for the influence of class and score on the assessment for learning (Hypothesis 1). This hypothesis was rejected. A two-way ANCOVA of group and class was used to investigate the effects on proportion correct of the summative assessment, controlling for the achievement on the assessment for learning. No significant effect of the feedback condition on student achievement regarding the summative assessment was found.

Even though no significant effects were found between one feedback condition and another, the results of this study do give a clear indication of the type of feedback students perceive to be most useful for learning. Student responses on the questionnaires indicate that students perceive KCR + EF (immediate and delayed) to be more useful for learning than KR only. Furthermore, the results suggests that students prefer immediate feedback to delayed feedback. From the results of the questionnaire, it can be concluded that students perceive immediate KCR + EF to be most useful for learning. Also, students have a more positive attitude towards feedback in a CBA when they receive KCR + EF rather than KR only.

The claims that are made with regard to the effects of immediate and delayed feedback vary widely (Mory, 2004). Even though no effects on the learning outcomes were found with regard to the effectiveness of immediate or delayed feedback, the results from the time log confirm that the timing of feedback is an important aspect to take into account. It was expected that students in Group 3 would spend less time reading the feedback than would students in Groups 1 and 2 (Hypothesis 2). This hypothesis was rejected. Students in Group 1 spent more time reading the feedback in the assessment for learning than did students in Group 3. No difference was found between Groups 2 and 3 concerning the time spent reading feedback. This outcome is remarkable because while the content of the feedback for Groups 1 and 2 was identical, the feedback for Group 3 was much shorter and less complex and would thus take less time to read. Only the timing of the feedback within Groups 1 and 2 differed. This outcome clearly suggests that students spent more time reading feedback when the feedback was delivered immediately than when the feedback was delivered with a delay. It could be that the time spent reading feedback was also influenced by the test length, since it is assumed that students have limited time that they are willing to invest in low-stakes assessments (Wise, 2006).

Additionally, the questionnaire provided information about students' feedback-reading behaviour. Students' responses on the questionnaire suggest that they paid more attention to feedback for incorrectly answered items than for correctly answered items. This outcome is in line with Timmers and Veldkamp's (2011) claim that students pay more attention to feedback when they answer an item incorrectly than when they answer an item correctly. Also, from a case study that included two groups of university students, Miller (2009) found that students prefer immediate feedback to delayed feedback.

With regard to the time spent reading feedback, it was expected that the student characteristics of motivation and attitude would be positively related to the time spent reading feedback (Hypothesis 3). This hypothesis was not rejected because a slightly positive significant relationship was found for motivation, and a moderately positive relationship was found for attitude.

Several reasons could explain this study's lack of clear outcomes with regard to feedback effects. First of all, the sample size was small, which resulted in the statistical tests having low power. Also, the moment in the learning process at which students were subjected to a CBA for learning could have affected their limited growth with regard to the learning outcomes. Since the assessment for learning was administered directly prior to the summative assessment, it can be assumed that students had already studied the subject matter thoroughly. Therefore, the gap between the current and goal knowledge was presumably small. In other words, they might not have needed (or felt the need) to receive feedback, which is a condition that has to be met in order for feedback to be effective (Stobart, 2008). This could also explain the limited amount of time students spent reading the feedback. As well, within this experiment, students did not have a chance to adapt their learning or to look up information in their study materials before the summative assessment was administered. In other words, we did not give the students much opportunity to learn. In addition, the time limit for the assessment could have affected students' willingness to read the feedback as well as their motivation to learn. This implies that Stobart's second and third conditions for feedback to be effective might not have been met, meaning that students did not have sufficient time to use the feedback and were not willing and able to use the feedback. Besides, the students who participated in this experiment were not used to taking CBAs. It might be the case, therefore, that students only paid limited attention to the feedback because they did not accept the CBA (Terzis & Economides, 2011). Furthermore, in the comments box, many students reported that they found it hard to concentrate during the entire assessment session. This might have negatively affected students' performance on the summative assessment.

In this study, we did not find an effect of feedback on students' learning outcomes. Indeed, in the literature, there is not much evidence available that feedback in CBAs leads to student performances that are more successful (e.g., Clariana & Lee, 2001; Corbalan, et al., 2009; Gordijn & Nijhof, 2002; Kopp et al., 2008). Many reasons can be thought of as to why researchers do not succeed in finding convincing evidence regarding the effectiveness of various feedback types. We doubt that there is one best way of providing feedback, given the interaction between student characteristics, task characteristics and feedback characteristics. This is in line with the findings of Hattie and Timperley (2007) and Shute (2008), who conclude that the literature provides inconsistent results with respect to different methods for providing feedback on students' learning outcomes. Also, in many studies that investigate the effects of feedback on students' learning outcomes, the time students spend reading feedback is not taken into consideration. The results of this study, however, suggest that time spent reading feedback varies widely depending on the different ways of delivering feedback as well as between students within one feedback condition. Therefore, it is recommended that future research take into account time spent reading feedback. Unfortunately, the time students spent reading the feedback was not available at the item level within this experiment. If this data were to become available, it would be possible to investigate the relationship between item difficulty, the ability of the student and time spent reading feedback. This type of analysis could lead to new insights into the effects of different feedback types and feedback timing on learning, especially between students with varying ability levels. These insights could be a starting point for combining assessments for learning with computerised adaptive testing.

A limitation of this study was that the assessment for learning and the summative assessment were not constructed from a calibrated item pool. Unfortunately, after administering the summative assessment, we had to reduce the test length from 30 to 11 items due to the multidimensionality of the assessment and poor quality of some of the items. Also, the summative assessment appeared to be more difficult than the assessment for learning. It might, therefore, be possible that there was an effect as a result of the feedback condition, but the summative assessment was not sensitive enough to measure this effect.

In future research, it is recommended that longer assessments of previously calibrated items be used in order to develop assessments that are more reliable. Application of the Spearman Brown prophesy formula predicts that Cronbach's alpha will be above .80 for a comparable 30-item summative assessment. Besides, it is recommended to use parallel forms of assessments to compare the results for both the assessment for learning and the summative assessment. In this way, the effects of different feedback conditions can be measured with more precision than was the case in this study.

Previous research has shown that the effects of various methods for providing feedback differ concerning varying levels of learning outcomes. This study did not distinguish between items that measured a specific level of learning outcomes because of the limited amount of items used in the assessments. Making a distinction between different levels of learning outcomes could lead to more insight into the conditions under which feedback is effective.

In future research, it is recommended that larger groups be used in order to increase the statistical power, and therefore the chance of finding significant effects of different feedback conditions. Also, future research should point out if students benefit from computer-based assessments for learning in the long run. Since in this study the summative assessment was administered immediately after the assessment for learning, only short-term learning effects could be measured in this experiment.

In this study, only the effect of written feedback was investigated. However, CBAs provide more opportunities for providing feedback than only text; for example, one could deliver or support feedback using pictures, video, or audio. The usefulness of these media depends on many variables, such as the subject matter as well as the age and education level of the students. For example, it is possible that students with low reading ability or dyslexia benefit more from feedback provided by audio than from feedback provided by text.

This study provides many possible options for further research. However, the present software available for CBA does not allow CBAs for learning to be developed to their full potential. Within this experiment, many roundabout ways had to be taken in order to investigate the effects of both immediate and delayed feedback. This led to restrictions—for example, navigating between the items in the assessment for learning was not possible. Therefore, in years to come, it is recommended that the software for CBA should adapt to the needs within the (research) field of education. This would also make it possible to investigate the effects of feedback using item types that are more complex (Wiliamson, Mislevy, & Bejar, 2006). In conclusion, more research is needed in order to investigate the effects of different methods for providing feedback on students' learning outcomes.

References

Assessment Reform Group (1999). *Assessment for learning: Beyond the black box.* Retrieved from http://assessmentreformgroup.files.wordpress.com/2012/01/beyond_blackbox.pdf

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research 13*, 111–127. doi:10.2190/9LMD-3U28-3A0G-FTQT

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213–238. doi:10.3102/00346543061002213

Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dory, Y., Ridgway, J., Wiesemes, R., & Nickmans, G. (2006). A learning integrated assessment system. *Educational Research Review, 1*, 61–67. doi:10.1016/j.edurev.2006.01.001

Clariana, R. B., & Lee, D. (2001). The effects of recognition and recall study tasks with feedback in a computer-based vocabulary lesson. *Educational Technology Research and Development, 49*, 23–36. doi:10.1007/BF02504913

Clariana, R. B., Wagner, D., & Murphy, L. C. R. (2000). Applying a connectionist description of feedback timing. *Educational Technology Research and Development, 48*(3), 5–21. doi:10.1007/BF02319855

Corbalan, G., Paas, F., & Cuypers, H. (2010). Computer-based feedback in linear algebra: Effects on transfer performance and motivation. *Computers & Education, 55*, 692–703. doi:10.1016/j.compedu.2010.03.002

Gordijn, J., & Nijhof, W. J. (2002). Effects of complex feedback on computer-assisted modular instruction. *Computers and Education, 39*, 183–200. doi:10.1016/S0360-1315(02)00025-8

Hattie, J., & Gan, M. (2011). Instruction based on feedback. In P. Alexander & R. E. Mayer (Eds.), *Handbook of research on learning and instruction* (pp. 249–271). New York, NY: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112. doi:10.3102/003465430298487

Keller, J. M. (1983). Motivational design of instruction. In C. M. Reigeluth (Ed.), *Instructional theories and models: An overview of their current status* (pp. 383–434). Hillsdale, NJ: Erlbaum.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284. doi:10.1037/0033-2909.119.2.254

Kopp, V., Stark, R., & Fischer, M. R. (2008). Fostering diagnostic knowledge through computer-supported, case-based worked examples: Effects of erroneous examples and feedback. *Medical Education, 42*, 823–829. doi:10.1111/j.1365-2923.2008.03122.x

Lee, H. W., Lim, K. Y., & Grabowski, B. L. (2010). Improving self-regulation, learning strategy use, and achievement with metacognitive feedback. *Educational Technology Research and Development, 58*, 629–648. doi:10.1007/s11423-010-9153-6

Lopez, L. (2009). *Effects of delayed and immediate feedback in the computer-based testing environment* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3358462)

Miller, T. (2009). *Formative computer-based assessments: The potentials and pitfalls of two formative computer-based assessments used in professional learning programs* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 305048958)

Morrison, G. R., Ross, S. M., Gopalakrishnan, M., & Casey, J. (1995). The effects of feedback and incentives on achievement in computer-based instruction. *Contemporary Educational Psychology, 20*, 32–50. doi:10.1006/ceps.1995.1002

Mory, E. H. (1994). Adaptive feedback in computer-based instruction: Effects of response certitude on performance, feedback-study time, and efficiency. *Journal of Educational Computing Research, 11*(3), 263–290. Retrieved from http://www.editlib.org/p/78600

Mory, E. H. (2004). Feedback research revisited. In D. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 745–783). Mahwah, NJ: Lawrence Erlbaum Associates.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153–189. doi:10.3102/0034654307313795

Smith, P. L., & Ragan, T. J. (2005). *Instructional design* (3rd ed.). New York, NY: Wiley.

Smits, M., Boon, J., Sluijsmans, D. M. A., & van Gog, T. (2008). Content and timing of feedback in a web-based learning environment: Effects on learning as a function of prior knowledge. *Interactive Learning Environments, 16*, 183–193. doi:10.1080/10494820701365952

Stobart, G. (2008). *Testing times: The uses and abuses of assessment.* Abingdon, England: Routledge.

Terzis, V., & Economides, V. V. (2011). The acceptance and use of computer based assessment. *Computers & Education, 56,* 1032–1044. doi:10.1016/j.compedu. 2010.11.017

TiaPlus (Version 2009) [Computer software]. Arnhem, the Netherlands: Cito.

Timmers, C. F., & Veldkamp, B. P. (2011). Attention paid to feedback provided by a computer-based assessment for learning on information literacy. *Computers & Education, 56,* 923–930. doi:10.1016/j.compedu.2010.11.007

Question Mark Perception (Version 4.0) [Computer software]. Available from http://www.questionmark.co.uk/us/index.aspx

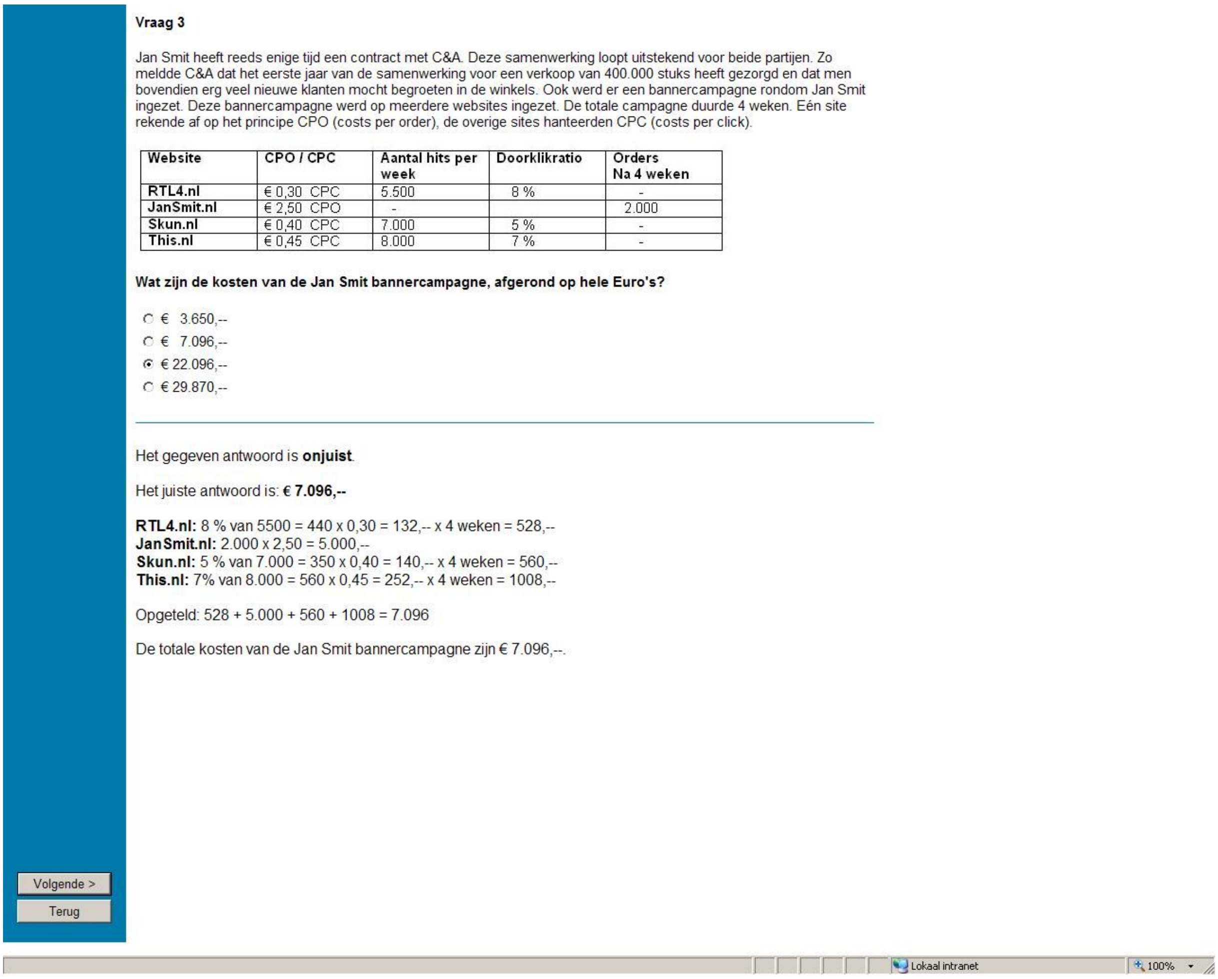
Verhelst, N. D., Glas, C. A. W. & Verstralen, H. H. F. M. (1995). OPLM: One Parameter Logistic Model (Version 3.0) [Computer software]. Arnhem, the Netherlands: Cito.

Wang, T. (2011). Implementation of web-based dynamic assessment in facilitating junior high school students to learn mathematics. *Computers & Education, 56,* 1062–1071. doi:10.1016/j.compedu.2010.09.014

Wiliamson, M. D., Mislevy, R. J., & Bejar, I. I. (2006). *Automated scoring of complex tasks in computer-based testing.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement In Education, 19*(2), 95–114. Retrieved from http://www.editlib.org/p/68431

Appendix 2A. Item 3 in the assessment for learning and KCR + EF



Vraag 3

Jan Smit heeft reeds enige tijd een contract met C&A. Deze samenwerking loopt uitstekend voor beide partijen. Zo meldde C&A dat het eerste jaar van de samenwerking voor een verkoop van 400.000 stuks heeft gezorgd en dat men bovendien erg veel nieuwe klanten mocht begroeten in de winkels. Ook werd er een bannercampagne rondom Jan Smit ingezet. Deze bannercampagne werd op meerdere websites ingezet. De totale campagne duurde 4 weken. Eén site rekende af op het principe CPO (costs per order), de overige sites hanteerden CPC (costs per click).

Website	CPO / CPC	Aantal hits per week	Doorklikratio	Orders Na 4 weken
RTL4.nl	€ 0,30 CPC	5.500	8 %	-
JanSmit.nl	€ 2,50 CPO	-		2.000
Skun.nl	€ 0,40 CPC	7.000	5 %	-
This.nl	€ 0,45 CPC	8.000	7 %	-

Wat zijn de kosten van de Jan Smit bannercampagne, afgerond op hele Euro's?

○ € 3.650,--
○ € 7.096,--
● € 22.096,--
○ € 29.870,--

Het gegeven antwoord is **onjuist**.

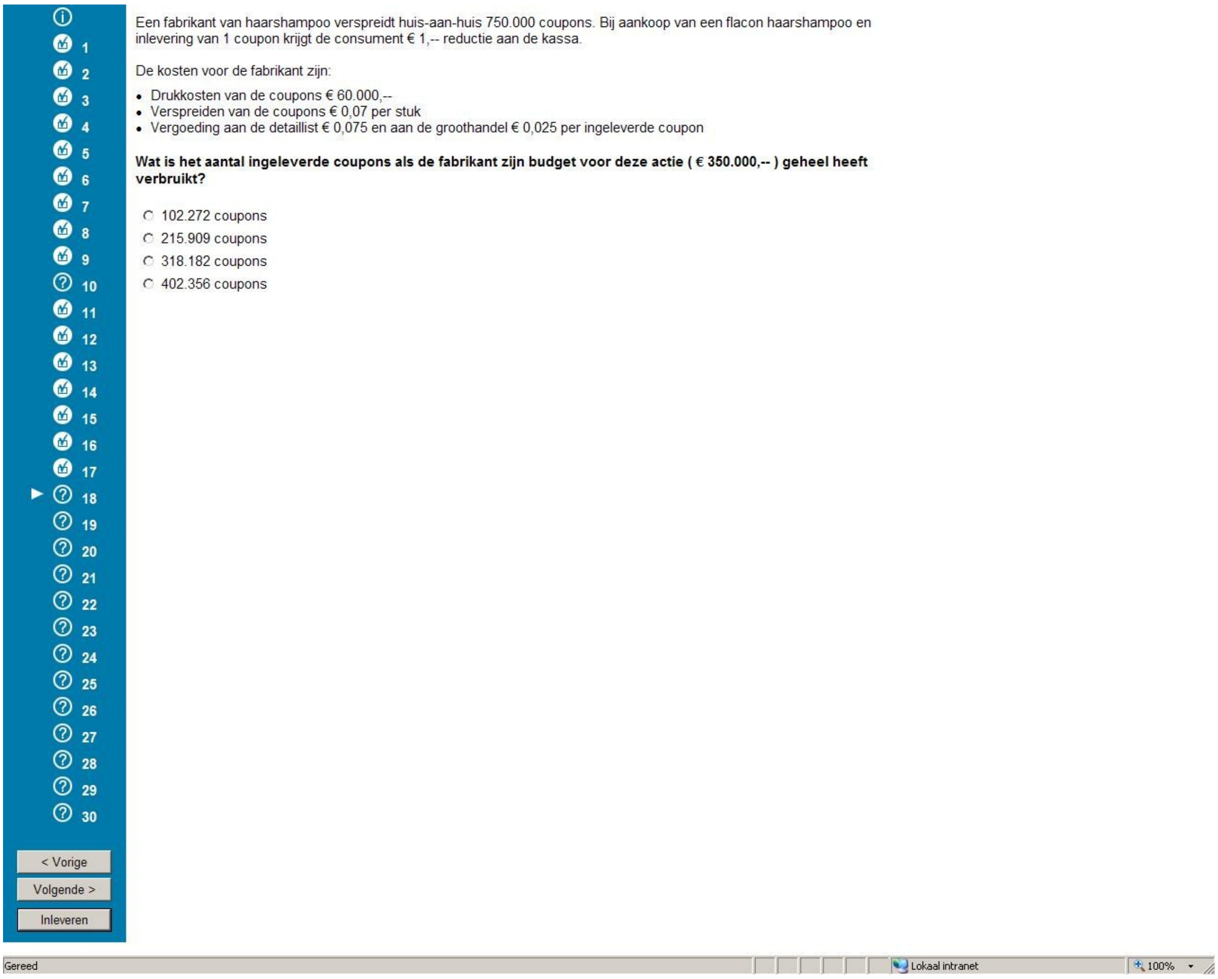
Het juiste antwoord is: **€ 7.096,--**

RTL4.nl: 8 % van 5500 = 440 x 0,30 = 132,-- x 4 weken = 528,--
JanSmit.nl: 2.000 x 2,50 = 5.000,--
Skun.nl: 5 % van 7.000 = 350 x 0,40 = 140,-- x 4 weken = 560,--
This.nl: 7% van 8.000 = 560 x 0,45 = 252,-- x 4 weken = 1008,--

Opgeteld: 528 + 5.000 + 560 + 1008 = 7.096

De totale kosten van de Jan Smit bannercampagne zijn € 7.096,--.

Appendix 2B. Item 18 in the summative assessment



Een fabrikant van haarshampoo verspreidt huis-aan-huis 750.000 coupons. Bij aankoop van een flacon haarshampoo en inlevering van 1 coupon krijgt de consument € 1,-- reductie aan de kassa.

De kosten voor de fabrikant zijn:

- Drukkosten van de coupons € 60.000,--
- Verspreiden van de coupons € 0,07 per stuk
- Vergoeding aan de detaillist € 0,075 en aan de groothandel € 0,025 per ingeleverde coupon

Wat is het aantal ingeleverde coupons als de fabrikant zijn budget voor deze actie (€ 350.000,--) geheel heeft verbruikt?

○ 102.272 coupons
○ 215.909 coupons
○ 318.182 coupons
○ 402.356 coupons

Appendix 2C. Item 4 and item 6 in the assessment for learning including KCR + EF

4. The price elasticity (Ev) of the demand for a certain product is -5. What does this mean?

a. It is a luxury product.
b. It is a substitute product.
c. There is an in-elastic demand for this product.
d. There is an elastic demand for this product.

Feedback:

The answer you provided is **correct / incorrect**.
The correct answer is: **There is an elastic demand for this product.**

The rule is: if Ev is smaller than -1, the demand for the product is elastic (price-sensible). In this case, there is an elastic demand for this product, since EV is -5.

6. In which of the following four situations can collective advertising be used effectively for a biscuit company?

a. In the situation the biscuit company introduces a new product.
b. In the situation that the competition on the market is intensified as a consequence of new providers of biscuits.
c. In the situation that the product of the biscuit company is the only brand-product in the product group wholemeal biscuits.
d. In the situation that the company and its competitors are faced with a declining turnover of biscuits.

Feedback:

The answer you provided is **correct / incorrect**.
The correct answer is: **In the situation that the company and its competitors are faced with a declining turnover of biscuits.**

Collective advertising is advertising by all companies in an industry.
In the situation a company and its competitors are faced with a declining turnover of a certain product, all companies within the biscuit industry have an interest in creating more demand for biscuits. An effective way to do so is to collectively advertise the product biscuits.

Chapter 3. The Effectiveness of Methods for Providing Written Feedback through a Computer-Based Assessment for Learning: A Systematic Review[3]

Abstract

This study reviews literature regarding the effectiveness of different methods for providing written feedback through a computer-based assessment for learning. In analysing the results, a distinction is made between lower-order and higher-order learning. What little high-quality research is available suggests that students could benefit from knowledge of correct response (KCR) to obtain lower-order learning outcomes. As well, elaborated feedback (EF) seems beneficial for gaining both lower-order and higher-order learning outcomes. Furthermore, this study shows that a number of variables should be taken into account when investigating the effects of feedback on learning outcomes. Implications for future research are discussed.

[3] This chapter is a minor revision of Van der Kleij, F. M., Timmers, C. F., & Eggen, T. J. H. M. (2011). The effectiveness of methods for providing written feedback through a computer-based assessment for learning: A systematic review. *CADMO, 19*, 21–39. doi:10.3280/CAD2011-001004

3.1. Introduction

The effects of feedback on learning have been investigated to a large extent (e.g., Hattie & Timperley, 2007; Mory, 2004; Shute, 2008). However, the literature conveys conflicting results (Kluger & DeNisi, 1996; Shute, 2008). More specifically, there is not much evidence concerning which feedback interventions might positively affect student learning in a computer-based environment. These environments make it possible to deliver individualized feedback in a timely manner, which is generally claimed to positively influence the learning process. Integrating assessment into the learning process is one of the main features of the assessment for learning approach (for more information, see Stobart, 2008). To date, no systematic reviews have been published that investigate the effectiveness of written feedback in a computer-based assessment (CBA) for learning. The purpose of this study is to investigate the effectiveness of different methods for providing feedback in a CBA as well as to identify gaps in current knowledge on this topic.

3.1.1 Computer-Based Assessment for Learning

Computer-based assessment (CBA) has increased in popularity in the preceding years. CBA has some advantages over traditional paper-and-pencil tests, such as higher test efficiency, automated scoring, and the possibility of adapting the item difficulty to the ability of the students. However, not only does CBA have practical advantages, the literature suggests that CBAs can have valuable pedagogical advantages when used for formative purposes. Namely, it is possible to provide the test taker with feedback through the computer. The way this feedback is provided can vary from simply telling the test taker that the answer is wrong to providing an extensive explanation regarding the learning content which is assessed by the item. Stobart (2008) states that in assessment for learning, feedback "is seen as the key to moving learning forward" (p. 144). Thus, CBA makes it possible to deliver feedback to students while they are taking the test. It is claimed that this has a positive effect on students' learning outcomes because their discrepancies between their current understanding and the goal can immediately be resolved (Hattie & Timperley, 2007).

3.1.2 Ways of Providing Feedback

Hattie and Timperley (2007) define feedback as "information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding" (p. 81). The primary aim of providing students with feedback is to close the gap between their current status in the learning process and the intended learning outcomes. Furthermore, feedback can provide students with insight into their own learning process and thereby support learning. However, there are many possible ways of providing feedback, not all of which have a positive effect on the learning outcomes of students (Stobart, 2008). In this study, we classified feedback based on types (Shute, 2008), levels (Hattie & Timperley, 2007), and timing (Shute, 2008).

Shute (2008) classified different types of feedback based on specificity, complexity, and length. For example, knowledge of results (KR) is a low-complex type of feedback: The student is merely told whether the answer is correct or incorrect. The correct answer is not provided. Knowledge of correct response (KCR) is a type of feedback in which the student is told the correct answer. Another type of feedback distinguished by Shute is termed elaborated feedback (EF). There is no clear definition for EF; the degree of elaboration differs widely. EF could, for example, mean that the student is presented with worked-out solution steps or is directed to study material. Besides these types of feedback, a distinction can be made between single-try and multiple-try feedback. This implies students can answer an item again after an initial incorrect response.

Hattie and Timperley (2007) distinguish four levels at which the feedback can be aimed—this is an expansion of a previously developed model by Kluger and DeNisi (1996). They distinguish between the self, task, process, and regulation levels. Feedback at the self level is not related to the task performed but is aimed at learner characteristics. Feedback at the task level is aimed at correcting work. Process-level feedback relates to the process that was followed to perform a particular task and gives suggestions regarding how the process can be improved. Feedback at the regulation level relates to processes in the mind of the learner, such as those of self-assessment, willingness to receive feedback and self-regulation in learning.

Also, the timing of feedback can differ. In the literature, immediate and delayed feedback is distinguished. Immediate feedback is usually delivered directly after the student has responded to an item (Shute, 2008). There is no univocal definition for delayed feedback because there is a wide range of possibilities for the degree of delay in which the feedback is delivered. In CBAs, delivering feedback often happens relatively quickly because the computer itself generates the feedback. In this study, the term *delayed feedback* is used for all feedback that is not delivered immediately after completing each item.

3.1.3 Feedback and Learning

This study's focus is on the effects of feedback on students' learning outcomes. *Learning outcomes* is a broad term that describes the outcomes of a learning process, one in which a student has executed particular tasks. Smith and Ragan (2005) emphasize that

> Some learning tasks are substantially different from others in terms of the amount and kind of cognitive effort required in learning, in the kinds of learning conditions that support their learning, and in the way to test for their achievement. (p. 79)

Also, they claim that different ways of providing feedback are differentially advantageous for certain levels of learning outcomes.

In this study, a distinction is made between lower-order and higher-order learning outcomes. For this purpose, two fundamental theories are combined, as described by Smith and Ragan (2005). These theories are Gagné's (1985) types of learning outcomes and Bloom, Englehart, Furst, Hill, and Krathwohl's (1956) taxonomy. Gagne's *declarative knowledge* and Bloom et al.'s *recall* and *understanding* are categorized as lower-order learning outcomes. These types of learning demand that students recall, recognize or understand something without the need to apply this knowledge.

In higher-order learning outcomes, an application of the knowledge gained is required. Gagné (1985) refers to these learning outcomes as *intellectual skills*. In the theory developed by Bloom et al. (1956), these types of learning are divided into *application*, *analysis*, *synthesis,* and *evaluation*. Here, students are required to apply their acquired knowledge in new situations, which is called transfer (Smith & Ragan, 2005). Smith and Ragan state that KR is most suitable for declarative knowledge, that KCR can facilitate both declarative knowledge and intellectual skills acquisition and that EF is the most appropriate for obtaining intellectual skills.

Many variables influence the relation between feedback and learning. Timmers and Veldkamp (2011) categorize these variables into characteristics of the task (e.g., task difficulty), characteristics of the feedback intervention (as discussed in paragraph 3.1.2), and characteristics of the learner (e.g., engagement in feedback). Their study shows that the extent to which students pay attention to the feedback from a CBA for learning can vary. Their focus depends, for example, on the length of the assessment and whether or not the student answered the item correctly. These findings show that students do not inevitably engage with feedback. This needs to be considered, since the mindful engagement of feedback is thought to be crucial for promoting learning (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991).

3.2. Method

3.2.1 Procedure

The procedure for this study is based on the method for executing systematic reviews in the social sciences as proposed by Petticrew and Roberts (2006). The review process includes three phases. First, the research question is defined, the databases and search terms are chosen, the literature search is carried out, the inclusion criteria are defined, and, eventually, the studies found in the literature search are selected using the inclusion criteria. Next, the characteristics of the selected studies are analysed using a data-extraction form. This implies that the same data are gathered from all the selected studies. Also, the quality of the selected studies is assessed in this phase in order to determine how much weight should be given to the study's findings. Subsequently, the conclusions of the different studies are brought together and synthesized using a qualitative method (narrative analysis). This implies that a systematic description of the study results is given in which the individual results of studies are presented schematically. These results are integrated in order to draw the final conclusion.

3.2.2 Databases and Search Terms

For the literature search,[4] four databases were used: ERIC, PsycInfo, Scopus, and Web of Science. These are the most commonly used databases for educational research. The search terms were determined by using a thesaurus and by analysing terminology used in references to relevant studies. The search terms were very specifically defined in order to obtain as many relevant hits as possible. Search terms included terms related to CBA or computer-based learning environments (e.g., computer-based assessment, computer-based instruction, computer-based formative assessment, etc.), and terms related to feedback (e.g., formative assessment). The exact search string can be found in Appendix 3A.

[4] The literature search was performed in November 2010

Search terms related to CBA were searched for in every field from each database. Search terms related to feedback were searched for in the title, abstract, or keyword fields. This is because if the main goal of the study is related to feedback, it can be assumed the term would be mentioned in at least one of these search fields. As well, the search in Scopus was restricted to the social sciences. Finally, in Web of Science, the search was narrowed down to terms related to CBA that appeared as topics as well as terms related to feedback that appeared in titles.

3.2.3 Inclusion Criteria

In order to select the relevant studies from the bulk of studies found, useful and clear inclusion criteria were needed. After a short field trial, the following inclusion criteria were chosen:

1. *The study has been published in the English language;*
2. *The study has been published in a journal article;*
3. *The dependent variable in the study is the learning outcomes of the students (achievement);*
4. *The main goal of the study is to compare the effects of different feedback interventions (of which at least one is written feedback) in a computer-based environment on the learning outcomes of individual students;*
5. *The study has been published in a journal that is listed in the Social Science Citations Index or Social Science Citations Index Expanded (Thomson Reuters, 2010).*

The first inclusion criterion aimed to select only studies that have a high degree of accessibility. If a study is accessible, this will increase the chance of making a contribution to the scientific knowledge within a certain field.

By using the second inclusion criterion, studies like reports, dissertations, or conference proceedings were removed from the selection. Regarding these kinds of documents, it is not clear if and how they have been reviewed. Therefore, it is unknown whether these studies live up to basic scientific standards. This step served to establish a rough initial selection of relevant types of studies.

The third inclusion criterion resulted in the selection of studies in which the learning outcomes of students is the dependent variable. Existing review studies were excluded in this study. However, some still contributed to creating the theoretical framework for analysing the results of this study (e.g., Shute, 2008).

By assessing the studies using the fourth criterion, studies with no written feedback were excluded from the selection. Also, studies that focus on feedback that is not provided within a computer-based environment were excluded. Studies that focus on feedback by students for teachers or on feedback from student to student were also excluded.

The fifth inclusion criterion was meant to select only high-quality studies that have been published in peer-reviewed, influential journals. If the study has been published in a journal that is not listed in the Social Science Citations Index (SSCI) or Social Science

Citations Index Expanded (SSCI/E), it was excluded. In a later stage of the process, the quality of the remaining studies was investigated more thoroughly.

3.2.4 Selection Process

All studies were exported to Thomson Reuters Endnote X4 (2010). With the assistance of this program, all duplicate studies were found and subsequently removed from the selection. Next, using the inclusion criteria, the studies were successively screened based on their relevance for this review study. Whenever a study did not comply with a certain criterion, it was removed from the selection. If it was not clear whether a study complied with a criterion, it was not excluded. During the selection process, studies were judged based on their title, keywords, and abstract. Selection steps 1, 2, 3, and 5 were performed by one researcher. In selection step 4, the selection of a quarter of the studies was done by two researchers, independently. The researchers agreed in 95% of the cases, and the few discrepancies were resolved. Of the studies that satisfied the inclusion criteria, their full-text documents were obtained.

3.2.5 Data Extraction and Appraisal of Studies

Data were extracted using a data-extraction form. For every article, a data-extraction form was filled in.[5] The first part of the form contains information about the study setting, method, respondents, types of feedback, levels of feedback, timing, conclusions, etc. The second part of the form is concerned with the quality of the study, which was judged based on a method proposed by Petticrew and Roberts (2006). All studies were judged on five categories: general orientation, selection of the sample, method, data and statistical tests, and conclusions. A score was assigned to every category (-, +-, or +, representing 0, 1, or 2 points respectively), and the scores for each study were aggregated into a sum score. Subsequently, the studies were classified as being of low (0–3.5), moderate (4–7) or high (7.5–10) quality.

Three researchers filled out the data-extraction forms independently of each other. One researcher extracted data from 100% of the articles, a second researcher extracted data from 35% of the articles and another researcher extracted data from another 39% of the articles. Eventually, discrepancies were discussed and resolved. For the judgments of quality, when there was more than one reviewer, the average of the judgements was taken.

3.3. Results

3.3.1 Search Results

The literature search resulted in 127 hits in ERIC, 134 hits in PsycInfo, 400 hits in Scopus, and 524 hits in Web of Science. The total amount of studies found in the search was 1185. After removing duplicates, 1158 studies remained. These studies were subjected to the inclusion criteria.

[5] An example of the data-extraction form can be found in Appendix 3B.

3.3.2 Selection Results

In the first and second selection steps, respectively 20 and 166 studies were excluded. In the third step—searching for learning outcomes as an independent variable—666 studies were removed from the selection. In step four, 265 studies were excluded. Finally, in the fifth selection step, another 19 studies were removed from the selection. The final selection contains 22 articles. The results of the selection process are summarised in Table 3.1.

Table 3.1
Results selection Steps 1to 5

Selection step	Studies subjected	Studies excluded	Percentage excluded	Studies selected
1	1158	20	2%	1138
2	1138	166	15%	972
3	972	666	69%	306
4	306	265	87%	41
5	41	19	46%	22

3.3.3 Data Analysis

The analysis of the content took place for 18 of the 22 selected articles because four studies were classified as being of low quality. Of these 18 studies (indicated with an asterisk * in the reference list), only nine reported that one feedback condition was significantly superior to another.[6] Two studies investigated the effects of a CBA either with or without feedback and a paper-and-pencil test (Wang, Wang, Wang, & Huang, 2006; Wang, 2007). Post-test results of students in a CBA that included feedback were significantly higher than the results of students in the other two conditions. However, there were six different types of feedback included in the CBA, which implies that no effects can be assigned to a specific type of feedback. Only seven out of 18 studies analysed provided us with usable significant results regarding a specific way of providing feedback. An overview of the feedback interventions in the different studies is provided in Table 3.2. This table also shows which studies reported significant positive effects. Subsequently, the studies with significant effects are discussed.

The feedback types distinguished by Shute (2008) as described in Section 3.1.2 are used to present the results of the selected studies in Table 3.2. In addition, additional feedback characteristics are presented, because the EF interventions differ from one another. The KR interventions were mostly combined with a try-again option. Also, the timing of feedback (Shute, 2008) was taken into consideration, as was the level of feedback (Hattie & Timperley, 2007) and the level of learning outcomes (Bloom et al., 1956; Gagné, 1985).

[6] For this study, the only methods considered were those whose feedback regarding student achievement produced significant effects. Within the selected studies, some other significant effects were found, such as the effects of different feedback types on student motivation or test anxiety.

Five of the studies reported significant results of feedback in relation to lower-order learning outcomes. Lee, Lim, and Grabowski (2010) found that in a computer-based learning environment, students who received EF in the form of metacognitive feedback (at the task and regulation levels) performed significantly better—in terms of both comprehension and recall—than students who received KR. The results of this study suggest that providing metacognitive feedback leads to enhanced student self-regulation, whereupon students tackle learning tasks more effectively, which eventually leads to higher achievement.

Clariana, Wagner, and Murphy (2000) found that delayed feedback led to the highest retention and recognition levels compared to immediate KR + try again and KCR. The effects were especially evident for difficult items.

Butler, Karpicke, and Roediger (2008) concluded that providing students with immediate KCR leads to better retention of both correct and incorrect responses compared to not providing feedback and not administering an initial test.

Murphy (2007) did not find a significant main effect of the delayed KCR and delayed EF + hints + try again feedback types. However, an interaction effect was found regarding the manner of study and feedback types. Students working in pairs scored significantly higher when receiving delayed EF + hints + try again. Additionally, students who worked individually scored significantly higher when receiving delayed KCR. A small footnote in this study is that the experiment was performed in Japan with what may have been highly motivated students. In order for the delayed EF + hints + try again method to work, students must be highly engaged in performing the learning tasks. The fact that no main effect was found for the type of feedback could be due to the possibility that students who worked individually were not sufficiently engaged. Perhaps the way students' work affects the attention they give to feedback and, thereby, the effect the feedback has on their learning outcomes. Additionally, it is possible that students who worked in pairs supported each other by means of processing the feedback mindfully.

Pridemore and Klein (1995) investigated the effect of immediate KCR, immediate EF (including KCR) and no feedback on items integrated into lesson material. Within this study, the group who received EF (including KCR) scored significantly higher than the group who received KCR only. However, the group who received no feedback also scored significantly higher than the KCR group. The authors give some possible explanations for these unexpected effects. For the KCR + EF group, the additional information in the feedback probably aided the students in remembering the learning content. Furthermore, they indicate it is possible that students who did not receive feedback were more engaged with the lesson materials and, therefore, studied them more thoroughly. On the contrary, students who received KCR were probably not motivated to make any additional effort because they already knew the correct answer.

The results in Table 3.2 suggest that providing students with KR or KCR is not beneficial for facilitating higher-order learning outcomes. The two studies reporting significant results of feedback in relation to higher-order learning outcomes were those that provided students with EF. Corbalan, Paas, and Cuypers (2010) found that students who received immediate EF (including KCR) during all solution steps performed significantly better on transfer tasks than students who received delayed KCR on the final solution step only. The effect was not present for retention tasks. These results suggest that providing students with immediate stepwise EF at the process level is beneficial for the transfer of learning.

Table 3.2

Results of 18 Studies for Different Methods for Providing Feedback

Type of feedback	**Additional feedback characteristics**	**Timing of feedback**	**Level(s) of feedback**	**Level(s) of learning (L / H)***	**Positive effect at $\alpha = 0.05$**	**First author and publication year**
KR	Try again	Immediate	Task	L	No	Clariana, 2000
KR	Try again	Immediate	Task	L	No	Clariana, 2001
KR	Note taking	Immediate	Task	L	No	Lee, 2010
KR	Try again	Immediate	Task	L + H	No	Morrison, 1995
KR	Try again	Immediate	Task	H	No	Park, 1992
KR	-	Immediate	Task	L	No	Wise, 1989
KR	Display of current score	Immediate	Task	L	No	Wise, 1989
KR	-	-	Task	H	No	Kopp, 2008
KCR	-	Immediate	Task	L	Yes	Butler, 2008
KCR	-	Immediate	Task	L	No	Clariana, 2000
KCR	-	Immediate	Task	L	No	Clariana, 2001
KCR	Overt response	Immediate	Task	L	No	Clariana, 2001
KCR	-	Immediate	Task	L	No	Gordijn, 2002
KCR	-	Immediate	Task	L + H	No	Morrison, 1995
KCR	-	Immediate	Task	L	No	Pridemore, 1995
KCR	-	Immediate	Task	H	No	Roos, 1997
KCR	-	Delayed	Task	L + H	No	Corbalan, 2010
KCR	-	Delayed	Task	L	Yes	Clariana, 2000
KCR	-	Delayed	Task	L + H	No	Morrison, 1995
KCR	-	Delayed	Task	L	Yes[a]	Murphy, 2007
KCR	-	-	Task	L + H	No	Schwartz, 1993
EF, KR	Explanations + try again	Immediate	Task and process	H	No	Park, 1992
EF, KR	Explanations and further consequences	-	Task and process	H	No	Kopp, 2008
EF, KCR	Worked out solutions	Immediate	Task and process	H	No	Corbalan, 2009
EF, KCR	Hints	Immediate	Task and process	L + H	Yes[b]	Corbalan, 2010
EF, KCR	Hints + try again	Immediate	Task and process	L	No	Gordijn, 2002
EF, KCR	Explanations	Immediate	Task and process	L	Yes[c]	Pridemore, 1995
EF, KCR	Solution steps	Immediate	Task and process	H	**Yes**[d]	Smits, 2008
EF, KCR	Worked out solutions and explanations	Immediate	Task and process	H	No	Smits, 2008
EF, KCR	Solutions steps	Delayed	Task and process	H	**Yes**[d]	Smits, 2008
EF, KCR	Worked out solutions and explanations	Delayed	Task and process	H	No	Smits, 2008
EF	Metacognitive feedback	Delayed	Task and regulation	L	**Yes**	Lee, 2010
EF	Hints + try again	Delayed	Process	L	Yes[e]	Murphy, 2007
EF	Bayesian feedback	-	Task and process	L + H	No	Schwartz, 1993
EF	Bayesian feedback + rules	-	Task and process	L + H	No	Schwartz, 1993
EF, KR	Combination of formative assessment strategies	-	Task, process, and regulation	H	Yes	Wang, 2006
EF, KR	Combination of formative assessment strategies	-	Task, process, and regulation	H	**Yes**	Wang, 2007

Note: Significant effects reported in high quality studies are boldfaced.

* L = lower-order, H = higher-order.

[a] Effect was reported for students working individually.

[b] Effect was reported for transfer tasks, not for retention tasks.

[c] Students in the KCR + EF group and students in the no feedback group scored higher than the KCR group.

[d] Effect was reported for high-achieving students only.

[e] Effect was reported for students working in pairs.

Smits, Boon, Sluijsmans, and Van Gog (2008) conclude that in a web-based learning environment, high-achieving students benefit more from global EF than from elaborate EF. Remarkably, students reported that they appreciated the elaborate feedback more. In this experiment, global feedback contained KCR and solution steps. Elaborate EF included KCR, worked-out solution steps and accompanying explanations. No effects were found for the low-achieving students or regarding feedback timing.

Previous research showed that there is no guarantee students will pay attention to feedback (e.g., Stobart, 2008; Timmers & Veldkamp, 2011). In the selected studies, researchers seemed to assume that students pay attention to the feedback that is presented. Only one study examined the time students spent on reviewing material for both correct and incorrect responses (Morrison, Ross, Gopalakrishnan, & Casey, 1995). They found that the amount of time students spend on reading feedback is influenced by incentives. As well, the group of students who spent the most time reading the feedback also performed highest. In the other studies, the amount of attention paid to feedback by students was not taken into consideration. However, in six studies, the time students spent completing the entire assessment was logged (Clariana & Lee, 2001; Corbalan, Kester, & van Merriënboer, 2009; Corbalan et al., 2010; Kopp, 2008; Lee et al., 2010; Park & Gittelman, 1992).

The majority of the studies were conducted at universities (*n* = 10). The rest were conducted at high schools (*n* = 5), in vocational education settings (*n* = 2), or at a college (*n* = 1). None of the studies was conducted in primary education settings. No relationship was found between the type of education and feedback effects.

3.4 Discussion

The purpose of this study was to gain insight into effective methods for providing written feedback in a CBA and to identify gaps in current knowledge of this topic. In analysing the results, a distinction was made between feedback types (Shute, 2008), levels (Hattie & Timperley, 2007), timing (Shute, 2008), and the level of learning outcomes (Bloom et al., 1956; Gagné, 1985).

Of the 1158 studies found in the literature search, only 18 satisfied the inclusion criteria and some additional quality criteria. These outcomes clearly suggest that there is not much high-quality research available yet on this topic. Furthermore, of the 18 studies analysed, only nine found one feedback condition to be significantly superior to another. As well, only seven studies provided usable results for this study: In the studies performed by Wang et al. (2006) and Wang (2007), no effects could be assigned to a specific type of feedback. These studies do suggest, however, that providing students with certain feedback interventions can be beneficial for learning outcomes.

A limitation of this study—and of systematic reviews in general—is the impossibility of including all the studies relevant to the research question. This is because by using specific search strategies, one makes selections that do not necessarily include all relevant studies. Also, in this study, strict requirements were established regarding the quality of acceptable studies.

This study does not contain enough data to provide evidence on the effects of different ways of providing written feedback. It does, however, give an indication of what might be promising results under certain conditions. Additionally, the selected studies were conducted at either a university, a high school, a college or a vocational education institute; therefore, the conclusions of this study may not apply to primary education settings.

The results of this study suggest that for lower-order learning outcomes (recognition or recall), KCR and EF can be beneficial. It seems that KR is not effective though, regardless of whether it is combined with a try again option. The effects of KCR are mixed. Two studies investigated the effects of delayed KCR on lower-order learning outcomes, and both found a positive significant effect. This suggests that students could benefit from delayed KCR. However, the results suggest that providing students with KR or KCR is not beneficial for facilitating higher-order learning outcomes. This is in line with the statement made by Smith and Ragan (2005)—that such ways of providing feedback are mainly advantageous for declarative knowledge acquisition.

The EF interventions are not completely comparable because there are many variations possible. Nevertheless, the EF interventions show mostly positive significant effects. Those effects are present for both lower-order and higher-order learning outcomes. It seems that EF that is aimed at both the task and process levels or the task and regulation levels is beneficial for student learning. The effects for timing are not uniform. The results suggest that students can benefit from both immediate and delayed EF. It is clear that differential effects of feedback play a role in lower-order and higher-order learning outcomes.

Only three out of 18 studies investigated the effects of feedback at the regulation level. They all found a significant positive effect, although the findings of Wang et al. (2006) and Wang (2007) cannot be assigned to one specific way of providing feedback. Feedback at the regulation level thus seems effective, which was already reported by Hattie and Timperley (2007). However, more research is needed regarding the effects of feedback at the regulation level on students' learning outcomes.

The results also suggest that it is useful to make a distinction between students with initially differing ability levels. Some studies have found that different ways of providing feedback affect students differently depending on their ability levels (e.g., Smits et al., 2008). This is in line with Shute's (2008) statement that the effects of feedback differ among students with varying ability levels.

There are various possible explanations for the fact that only a few studies found that one feedback condition was favoured over another. First, the sample sizes of most selected studies were remarkably small, especially when considering the large amount of experimental groups involved. This resulted in the statistical tests having low power when comparing student achievement between different groups and made it improbable that significant results would be found. Also, the assessments used in most experiments contained very few items. In order to measure student learning with sufficient reliability, the number of items in the post-test should be satisfactory. Using high-quality instruments makes it possible to measure with high precision. In the majority of the studies, a post-test/summative assessment was administered directly after the assessment for learning. This implies students did not have much time to increase in ability. Therefore, we cannot expect huge effects over such a small time span. In order to find feedback effects, highly reliable assessments are needed. Additionally, only half of the studies reported information about the quality of the assessment(s) used in the experiment. Others did not report on the reliability of the assessment(s) at all. As well, the relationship between the underlying constructs of the sometimes-used pre-test, the assessment for learning, and the post-test/summative assessment was often unclear.

It is important to take into account the difficulty of the tasks in relation to the effects of feedback. Half of the studies provide information on the difficulty of the assessments used, of which six reported exact values (e.g., proportion correct). When looking at the results of those studies, it is striking that significant effects concerning a given method of providing feedback are found when using relatively difficult items (e.g., Butler et al., 2008).

Shute (2008) reviewed literature on formative feedback. She concluded that “there is no best way to provide feedback for all learners and learning outcomes” (p. 182). Given this, she developed guidelines for generating formative feedback, taking into account variables such as task difficulty. The results of this study suggest that when one wants students to learn from feedback, relatively difficult items are generally most suitable. This, in itself, is logical, since we can assume that when items that are more difficult are used, many will answer the items incorrectly, which implies a gap between the current and intended levels of the student. As a consequence, feedback provides an opportunity to fill this gap—for example, by clarifying misunderstandings in the learning process (Mory, 2004). However, if the items are relatively easy, with many students answering them correctly, only a few students can learn from the feedback because the others have already (almost) achieved the intended learning outcomes. In other words, using relatively difficult items provides opportunities to learn from feedback. Additionally, it is recommended that students’ previous ability levels be taken into account when analysing the assessment results. As well, determining the time students spent on reading the feedback can provide researchers with valuable information when used as a measure of students’ engagement with feedback.

In future research on the effects of feedback on learning, it is recommended that larger experimental groups be used. This is because larger groups will increase statistical power and, thus, the chance of finding significant effects. Furthermore, the characteristics of the task, the feedback intervention, and the learner should be taken into account in order to learn more about the effects of feedback on higher-order and lower-order learning outcomes.

References

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213–238. doi:10.3102/00346543061002213

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I, cognitive domain.* New York, NY: McKay.

* Butler, A. C., Karpicke, J. D., & Roediger III, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning Memory and Cognition*, *34*, 918–928. doi:10.1037/0278-7393.34.4.918

* Clariana, R. B., & Lee, D. (2001). The effects of recognition and recall study tasks with feedback in a computer-based vocabulary lesson. *Educational Technology Research and Development, 49*, 23–36. doi:10.1007/BF02504913

* Clariana, R. B., Wagner, D., & Murphy, L. C. R. (2000). Applying a connectionist description of feedback timing. *Educational Technology Research and Development, 48*, 5–21. doi:10.1007/BF02319855

* Corbalan, G., Kester, L., & van Merriënboer, J. J. G. (2009). Dynamic task selection: Effects of feedback and learner control on efficiency and motivation. *Learning and Instruction, 19*, 455–465. doi:10.1016/j.learninstruc.2008.07.002

* Corbalan, G., Paas, F., & Cuypers, H. (2010). Computer-based feedback in linear algebra: Effects on transfer performance and motivation. *Computers and Education, 55*, 692–703. doi:10.1016/j.compedu.2010.03.002

Gagné, R. M. (1985). *The conditions of learning* (4th ed.). New York, NY: Holt, Rinehart & Winston.

* Gordijn, J., & Nijhof, W. J. (2002). Effects of complex feedback on computer-assisted modular instruction. *Computers and Education, 39*, 183–200. doi:10.1016/S0360-1315(02)00025-8

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112. doi:10.3102/003465430298487

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284. doi:10.1037/0033-2909.119.2.254

* Kopp, V., Stark, R., & Fischer, M. R. (2008). Fostering diagnostic knowledge through computer-supported, case-based worked examples: Effects of erroneous examples and feedback. *Medical Education, 42*, 823–829. doi:10.1111/j.1365-2923.2008.03122.x

* Lee, H. W., Lim, K. Y., & Grabowski, B. L. (2010). Improving self-regulation, learning strategy use, and achievement with metacognitive feedback. *Educational Technology Research and Development, 58*, 629–648. doi:10.1007/s11423-010-9153-6

*Morrison, G. R., Ross, S. M., Gopalakrishnan, M., & Casey, J. (1995). The effects of feedback and incentives on achievement in computer-based instruction. *Contemporary Educational Psychology, 20*, 32–50. doi:10.1006/ceps.1995.1002

Mory, E. H. (2004). Feedback research revisited. In D. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 745–783). Mahwah, NJ: Lawrence Erlbaum Associates.

* Murphy, P. (2007). Reading comprehension exercises online: The effects of feedback, proficiency and interaction. *Language Learning and Technology, 11*(3), 107–129. Retrieved from http://llt.msu.edu/vol11num3/murphy/default.html

* Park, O. C, & Gittelman, S. S. (1992). Selective use of animation and feedback in computer-based instruction. *Educational Technology Research and Development*, *40*, 27–38. doi:10.1007/bf02296897

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide.* Oxford, UK: Blackwell.

* Pridemore, D. R., & Klein, J. D. (1995). Control of practice and level of feedback in computer-based instruction. *Contemporary Educational Psychology, 20*, 444–450. doi:10.1006/ceps.1995.1030

* Roos, L. L., Wise, S. L., & Plake, B. S. (1997). The role of item feedback in self-adapted testing. *Educational and Psychological Measurement, 57*, 85–98. doi:10.1177/0013164497057001005

* Schwartz, S., & Griffin, T. (1993). Comparing different types of performance feedback and computer-based instruction in teaching medical students how to diagnose acute abdominal pain. *Academic Medicine*, *68*, 862–864. doi:10.1097/00001888-199311000-00018

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153–189. doi:10.3102/0034654307313795

Smith, P. L., & Ragan, T. J. (2005). *Instructional design* (3rd ed.). New York, NY: Wiley.

* Smits, M., Boon, J., Sluijsmans, D. M. A., & Van Gog, T. (2008). Content and timing of feedback in a web-based learning environment: Effects on learning as a function of prior knowledge. *Interactive Learning Environments, 16*, 183–193. doi:10.1080/10494820701365952

Thomson Reuters (2010). Thomson Reuters EndNote (Version X4.0). [Computer software]. Available from: http://www.endnote.com/

Timmers, C. F., & Veldkamp, B. P. (2011). Attention paid to feedback provided by a computer-based assessment for learning on information literacy. *Computers & Education, 56,* 923–930. doi:10.1016/j.compedu.2010.11.007

* Wang, K. H., Wang, T. H., Wang, W. L., & Huang, S. C. (2006). Learning styles and formative assessment strategy: enhancing student achievement in Web-based learning. *Journal of Computer Assisted Learning*, *22*, 207–217. doi:10.1111/j.1365-2729.2006.00166.x

* Wang, T. H. (2007). What strategies are effective for formative assessment in an e-learning environment? *Journal of Computer Assisted Learning, 23*, 171–186. doi:10.1111/j.1365-2729.2006.00211.x

* Wise, S. L., Plake, B. S., Pozehl, B. J., & Barnes, L. B. (1989). Providing item feedback in computer-based tests: Effects of initial success and failure. *Educational and Psychological Measurement*, *49*, 479–486. doi:10.1177/0013164489492021

Appendix 3A. Search string

1) cbt OR cba OR “computer-based assessment” OR “computer-based assessments” OR “computer-based test” OR “computer-based tests” OR “computer-based testing” OR e-assessment OR “computer-based learning” OR “computer-based instruction” OR “computerized assessment” OR “computer-based formative assessment”

AND

2) feedback OR “formative evaluation” OR “formative assessment”

Appendix 3B. Data extraction form

Data to be extracted	**Notes from the researcher**
Title of the study	
Author(s)	
Year of publication	
Country/ countries in which the study was performed	
Journal	
Education level or stream (+ grade level)	
Sample size	
Subject	
The research goal as described by the authors	
Research question	
Statistical technique(s) for analysing the study results	
Dependent variable	
Software used for CBA	
Feedback conditions compared in study	
Item types in formative CBA	
Were the students randomly assigned to the treatment groups?	
Did the students complete the assessment in a supervised environment?	
Did the students complete the assessments voluntarily?	
Did the post-test contribute to the final grade of the subject?	
Did the authors report anything about the quality of the assessments?	
Did the authors report anything about the difficulty of the assessments?	
Did the authors take into account how much attention students paid to the feedback	
Did the learner have control in selecting the items in the assessments?	

Did the authors measure task-specific motivation (self-efficacy, goal orientation, task value)	
What is the time span between the intervention and post-test?	
Major conclusion	
General orientation 1. Which research question is the author trying to answer? 2. Is the research goal clearly defined? 3. Is the research in combination with the chosen method capable of finding a clear answer to the research question?	1. 2. 3. **Judgement:**
Selection sample 1. What is the sample size? 2. Does the study contain enough data to assure the validity of the conclusions? 3. Is it clear which sub group was involved in the study? 4. Is it clear in which country or countries the study was performed? 5. Did the authors indicate what the response rate was?	1. 2. 3. 4. 5. **Judgement:**
Method 1. Do the authors mention the statistical methods used? 2. Do the authors give an argumentation for the methods chosen? 3. Do the authors examine the relation between what they intend to measure and other variables? 4. Is the content of the different conditions clearly defined?	1. 2. 3. 4. **Judgement:**
Data and statistical tests 1. Were the data analysis performed in a precise and adequate way? 2. Do the authors mention the level of significance? 3. Do the authors present the results clearly?	1. 2. 3. **Judgement:**
Conclusions 1. Do the results resemble the conclusions of the research? 2. Do the authors present the conclusions clearly? 3. Do the researchers mention the limitations of their research?	1. 2. 3. **Judgement:**
Quality	Score:
Other remarks	

Chapter 4. Effects of Feedback in a Computer-Based Learning Environment on Students’ Learning Outcomes: A Meta-analysis[7]

Abstract

This meta-analysis investigated the effects of methods for providing item-based feedback in a computer-based environment on students’ learning outcomes. From 40 studies, 70 effect sizes were computed, which ranged from -0.78 to 2.29. A mixed model was used for the data analysis. The results showed that elaborated feedback (EF), e.g., providing an explanation, produced higher effect sizes (0.49) than feedback regarding the correctness of the answer (KR; 0.05) or providing the correct answer (KCR; 0.32). EF was particularly more effective than KR and KCR for higher-order learning outcomes. Effect sizes were positively affected by the feedback type EF. Larger effect sizes were found for mathematics compared to social sciences, science, and languages. Effect sizes were negatively affected by delayed feedback timing, and primary and high school. Although the results suggested that immediate feedback was more effective for lower-order learning than delayed feedback, and vice versa, no significant interaction was found.

[7] This chapter has been submitted as Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (submitted). *Effects of feedback in a computer-based learning environment on students’ learning outcomes: A meta-analysis.* Manuscript submitted for publication.

4.1 Introduction

The importance of assessment in the learning process is widely acknowledged, especially with the growing popularity of the assessment for learning approach (Assessment Reform Group [ARG], 1999; Stobart, 2008). The role of assessment in the learning process is crucial. “It is only through assessment that we can find out whether a particular sequence of instructional activities has resulted in the intended learning outcomes” (Wiliam, 2011, p .3). Many researchers currently claim formative assessment can have a positive effect on the learning outcomes of students. However, these claims are not very well grounded, an issue that has recently been addressed in detail by Bennett (2011), who argued that “the magnitude of commonly made quantitative claims for effectiveness is suspect, deriving from untraceable, flawed, dated, or unpublished resources” (p. 5). For example, the source that is most widely cited with regard to the effects of formative assessment is Black and Wiliam’s (1998a, 1998b, 1998c) collection of papers. Often, effect sizes of between 0.4 and 0.7 were cited from these studies, which suggests formative assessment had large positive effects on student achievement. Bennett argued, however, that the studies involved in their meta-analysis are too diverse to be expressed in a meaningful overall result. Consequently, an overall effect size for formative assessment is not very informative. Moreover, the meta-analysis itself has never been published and therefore could not be criticized. Bennett (2011) called the effect sizes in Black’s and Wiliam’s studies (1998a, 1998b) “a mischaracterization that has essentially become the educational equivalent of urban legend” (p. 12). Additionally, Bennett argued that other meta-analyses on formative assessment (e.g., Bloom, 1984; Nyquist, 2003; Rodriguez, 2004) have limitations and do not provide strong evidence. The most recently published meta-analysis on formative assessment (Kingston & Nash, 2011) has also already been criticised for its methodological aspects (Briggs, Ruiz-Primo, Furtak, Shepard, & Yin, 2012).

For meta-analyses to produce meaningful results, they have to focus on a specific topic in order to include studies that are sufficiently comparable. A key element of the assessment for learning approach is the feedback provided to students (ARG, 1999; Stobart, 2008). Various meta-analyses and systematic review studies have focused on the effects of feedback on learning outcomes (e.g., Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Kluger & DeNisi, 1996; Shute, 2008). The outcomes of these studies have not been univocal and were sometimes even contradictory. Many researchers have noted that the literature on the effects of feedback on learning provides conflicting results (e.g., Kluger & DeNisi, 1996; Shute, 2008). Given the current state of research, there is need for an updated meta-analysis focusing on specific aspects of formative assessment. The present meta-analysis concentrated on the effects of feedback provided to students in a computer-based environment.

4.1.1 Methods for Providing Feedback

In order to compare the effects of various methods of providing feedback, a clear classification of these methods is needed. In the current study, feedback was classified based on types (Shute, 2008), levels (Hattie & Timperley, 2007), and timing (Shute, 2008) as proposed by Van der Kleij, Timmers, and Eggen (2011). The focus of this study was item-based feedback. This feedback relates specifically to a student’s response to an item on a test.

Shute (2008) distinguished various types of feedback, which can be classified as knowledge of results (KR), knowledge of correct response (KCR), and elaborated feedback (EF). KR indicates whether the answer is correct or incorrect but does not provide the correct answer or any additional information. KCR is similar to KR, except the correct answer is provided. KR and KCR merely have a corrective function. According to Shute, EF can take many forms, such as hints, additional information, additional study material, or an explanation of the correct answer. One type of feedback, *error flagging*, which shows the student the location of the error but does not provide the correct answer or any additional information, can be classified as KR. Since the term EF has a wide range of possible meanings, the degree to which EF is effective for learning purposes varies widely. EF is often accompanied by KCR or KR, either implicitly or explicitly.

Shute (2008) classified *try again* as a feedback type. We, however, found it more useful to define this response as an additional feedback characteristic that could be combined with multiple types of feedback. It could, for example, be combined with KR or with KR and EF in the form of hints.

Hattie and Timperley (2007) used a model based on a study by Kluger and DeNisi (1996) to distinguish four levels of feedback. Feedback can be aimed at the self, task, process, and regulation levels. Feedback that is aimed at the self level does not relate to the task performed but instead relates to characteristics of the learner. An obvious example of feedback at the self level is praise; "You are a fantastic student!" Feedback at the task level serves a corrective function. Feedback at the process level addresses the process that has been followed to complete the task. Regulation level feedback is related to students' self-regulation; aspects that play a role here include self-assessment and willingness to receive feedback.

The feedback types distinguished by Shute (2008) are all task-related feedback types, which in this context means the feedback is item-specific, as opposed to, for example, summary feedback covering the entire assessment. Shute calls this *task level* feedback, but the task level outlined by Shute should not be confused with the task level feedback described by Hattie and Timperley (2007). In this study, whenever the word *task level* is used, it refers to Hattie and Timperley's definition.

Shute (2008) points out that with regard to timing, the results in the literature are conflicting even though the topic has been widely studied. With respect to feedback timing, a distinction can be made between *immediate* and *delayed* feedback. The definitions of immediate and delayed feedback seem to differ widely. This difference might be one of the reasons for the varying effects reported for immediate and delayed feedback (Mory, 2004). In formative assessment situations, immediate feedback is usually delivered right after a student has responded to an item. It is, however, hard to clearly define delayed feedback because of the wide variation in the possible degrees of delay (Shute, 2008). In computer-based environments, it is possible to provide students with feedback very quickly since the feedback is automatically given based on the student's response. Therefore, when it comes to feedback in a computer-based environment, delayed feedback can be defined unambiguously as "all feedback that is not delivered immediately after completing each item" (Van der Kleij et al., 2011, p. 23).

Recent research suggests that students prefer immediate to delayed feedback (Miller, 2009). An experiment performed by Van der Kleij, Eggen, Timmers, and Veldkamp (2012) showed that students spent significantly more time reading immediate feedback than delayed feedback. These results indicate that feedback timing is a relevant aspect to take into account when investigating the effects of feedback on learning.

4.1.2 Evidence on the Effects of Feedback in Computer-Based Environments

According to Bloom (1984), one-to-one tutoring is the most effective form of instruction. One-to-one tutoring is effective because the teacher can immediately intervene—provide feedback—when there is a misunderstanding. Therefore, instruction can continuously be adapted to the needs of the learner. Unfortunately, this type of instruction is unimaginable in today's educational systems. Current technology, however, offers promising solutions to this problem. Namely, if an assessment is administered in a computerized environment, it is possible to provide the student with standardised feedback based on his or her response to an item. Computer-based assessments (CBAs) have various advantages, such as the possibility of providing more timely feedback, automated scoring, and higher test efficiency (Van der Kleij et al., 2011). As in one-to-one tutoring situations, feedback in CBA can serve to immediately resolve the gap between students' current status in the learning process and the intended learning outcomes (Hattie & Timperley, 2007). In this way, CBA could assist teachers in providing students with individualized feedback. Similarly, most computer-based instruction (CBI) environments include practice questions with feedback.

To the best of our knowledge, only three studies have been published to date that provided an overview of the effects of feedback in computer-based environments on students' learning outcomes (Azevedo & Bernard, 1995; Jaehnig & Miller, 2007; Van der Kleij et al., 2011). An overview of the characteristics and main results of these studies will follow in chronological order.

Azevedo and Bernard (1995) conducted a meta-analysis to determine the effects of feedback on learning from CBI. In total, this meta-analysis included 22 studies (published between 1969 and 1992), which were mostly aimed at lower-order learning outcomes (see Smith & Ragan, 2005). The effect sizes from the studies ranged from 0.03 to 2.12. Azevedo and Bernard made a distinction between studies with an immediate post-test and studies with a delayed post-test. In total, they extracted 34 effect sizes from 22 studies that used an immediate post-test. The unweighted mean effect size was estimated at 0.80. A mean effect size of 0.35 was found for a delayed post-test. In analysing the data, the authors did not take any feedback characteristics (such as type) into consideration. It is only in the discussion section that the authors provided information on the characteristics of the feedback in the studies, which seemed to differ between the studies with an immediate post-test and those with a delayed post-test. This section of their study was, however, not very clarifying because the differences between the effect sizes found have not been related to specific feedback characteristics. Owing to the small sample, the datedness of the studies included, and the lack of identifying moderating variables, one could question the value of this meta-analysis for providing insight into what works in current educational practice.

Jaehnig and Miller (2007) conducted a systematic review of the effects of different feedback types in programmed instruction. As can be expected from the topic of this review study, many outdated studies were included. The publication years of the reviewed studies ranged from 1964 to 2004 ($N = 33$). In these studies, only the effects of feedback on lower-order learning outcomes, mostly recall, were investigated. Jaehnig and Miller concluded that KR is not effective for learning, that KCR is sometimes effective, and that EF seems to be most effective. EF in this study included any feedback that contained information in addition to information regarding the correctness of the response, such as an explanation. Furthermore, EF in this study also included answer until correct [AUC], a form of KR in which the student has to try again until the answer is correct. With regard to feedback timing, Jaehnig and Miller concluded that there were no differences between the effects of immediate and delayed feedback. However, in their study, no clear definitions of feedback immediacy and delay were provided. Sometimes delayed feedback was described as a certain number of seconds after the stimulus was presented to the student (e.g., 15 or 30 seconds after responding to the item), while other times it was delayed until after all items were completed. Furthermore, it is questionable to what degree the results of the studies included in their systematic review can be generalized to current educational practices owing to the large number of outdated studies and the different approach to learning.

Recently, a systematic review study was conducted by Van der Kleij et al. (2011). In their study, the effects found in various experiments on the effects of written feedback in a CBA were compared. Of the 18 studies selected for the review (published between 1989 and 2010), only 9 reported a positive effect of one feedback condition favouring another. One possible explanation for this is that the sample sizes, and therefore the statistical power, in these studies was generally small. Van der Kleij et al. concluded that KR seems ineffective, KCR seems moderately beneficial for obtaining lower-order learning outcomes, and EF seems to be beneficial for obtaining both lower-order and higher-order learning outcomes. The outcomes of their review study suggest that it is necessary to take the level of learning outcomes (Smith & Ragan, 2005) into account when examining the effects of feedback.

4.1.3 Objectives of the Present Study

The objective of this meta-analysis was to gain insight into the effectiveness of various methods for providing item-based feedback in a computer-based environment on students' learning outcomes. Conducting a meta-analysis makes it possible to detect patterns or effects that are not visible at the level of individual experiments. It also provides us with insights into the magnitude of the feedback effects.

It was not the aim of this meta-analysis to obtain an overall effect size expressing the effect of feedback in computer-based environments in general. In order to produce meaningful results, this meta-analysis had to provide multiple effect sizes: One for each type of feedback. The level of learning outcomes was also taken into account, which has been shown to be a relevant variable when examining feedback effects in a computer-based environment (Van der Kleij et al., 2011). It is unlikely that there is one ideal situation that has a positive influence on the learning outcomes of all students in all subjects. However, a meta-analysis offers the opportunity to reveal some effects, which were not present in the primary studies, which is especially relevant because the sample sizes in the primary studies are often small.

The present meta-analysis was built on the knowledge on the effects of feedback available to date. The question that is central to this meta-analysis is as follows: To what extent do various methods for providing item-based feedback in a computer-based learning environment affect students' learning outcomes?

Currently available evidence from the literature was used to construct hypotheses for this meta-analysis. Although the scope of this meta-analysis was narrower than that of those that have focused on feedback effects in classroom settings, the findings from the literature on feedback in classroom situations can provide usable insights as well. It was not the aim of this study to review the review studies on feedback effects. There are, however, some interesting findings available in the literature that can be used as a starting point of investigation with regard to the effects of feedback in computer-based environments:

- Feedback at the self-level (praise) has been shown to be ineffective (Kluger & DeNisi, 1996; Hattie & Gan, 2011; Hattie & Timperley, 2007).
- Simple feedback (that is only related to the correctness of the response) seems to be mainly effective for lower-order learning outcomes (Kluger & DeNisi, 1996; Van der Kleij et al., 2011).
- EF (which can take many forms) seems to be the most effective feedback type (Mory, 2004; Shute, 2008; Van der Kleij et al., 2011). However, due to the variations in the nature of EF, there is also a wide variation in the effects (Narciss, 2008; Shute, 2008).
- How the feedback is received differs from student to student (Hattie & Gan, 2011; Kluger & DeNisi, 1996; Stobart, 2008; Timmers & Veldkamp, 2011), and feedback has to be processed mindfully in order to have an effect on learning outcomes (Bangert-Drowns et al., 1991). Especially in a computer-based environment, students can easily ignore written feedback (e.g., Timmers & Veldkamp, 2011; Van der Kleij et al., 2012).
- The effectiveness of immediate versus delayed feedback seems to differ depending on the level of the intended learning outcomes (Shute, 2008). Shute suggested that when the feedback is intended to facilitate lower-order learning outcomes, immediate feedback works best, and when higher-order learning outcomes are at stake, it is best to provide feedback with a delay. However, the literature shows highly conflicting results when it comes to the timing of feedback.

The hypotheses of the current study are as follows:

1) KR and KCR have a small to moderately positive effect on lower-order learning outcomes.
2) KR and KCR have virtually no effect on higher-order learning outcomes.
3) EF has a moderate to large positive effect on lower-order learning outcomes.
4) EF has a moderate to large effect on higher-order learning outcomes.
5) There is an interaction effect between feedback timing and the level of learning outcomes. Immediate feedback is more effective for lower-order learning outcomes than delayed feedback and vice versa.

As benchmarks for the values of the effect sizes, we used Hattie's (2009) interpretation for the magnitude of effect sizes because they have been derived from the context of education. According to Hattie, an effect size of 0.2 can be considered small, an effect size of 0.4 can be considered moderate, and effect sizes of 0.6 are classified as large. Hypothesis 1 will be rejected when the effects of KR or KCR on lower-order learning outcomes are significantly lower than 0.2 or higher than 0.6. We will reject Hypothesis 2 when the effects of KR or KCR on higher-order learning outcomes are significantly larger than 0.2. Hypothesis 3 will be rejected when the effects of EF on lower-order learning outcomes are significantly below 0.4. When the effects of EF on higher-order learning outcomes are significantly below 0.4, Hypothesis 4 will be rejected. Hypothesis 5 will be discarded when there is no interaction effect between feedback timing and the level of learning outcomes, or when the expected direction of the effects is found to be reverse.

Furthermore, the relationships of various variables that seem relevant given the literature were explored, such as subject (e.g., Kingston & Nash, 2011), education level, level of learning outcomes (Van der Kleij et al., 2011), and the level at which the feedback is aimed (Hattie & Gan, 2011; Hattie & Timperley, 2007; Van der Kleij et al., 2011).

4.2 Method

4.2.1 Data Collection

For the data collection, a thorough and systematic search was conducted, which consisted of three phases. The primary search included the online databases ERIC, PsycInfo, ScienceDirect, Scopus, and Web of Science. The search was carried out using terms related to feedback, formative assessment, or assessment for learning in the title. Furthermore, the word *computer* or variations on this word had to appear in the abstract. In Web of Science, the search was restricted to terms related to *computer* in the title. In Scopus, the search was limited to the social sciences. No restrictions were made regarding the publication year and publication type during the search.

The references retrieved were exported to Endnote version X4 (Thomson Reuters, 2010) and assessed on their relevance using the inclusion criteria (which are specified in Section 4.2.2). Studies that met the inclusion criteria were retrieved in their full text forms. Subsequently, the *ancestry approach* (White, 1994) was used as a secondary method for searching studies, which means the reference section of each selected study was scanned for possibly relevant references for inclusion. Studies that met the criteria were also included. Third, the reference sections of existing meta-analyses and review studies on this topic (Azevedo & Bernard, 1995; Jaehnig & Miller, 2007; Van der Kleij et al., 2011) were scanned for possibly relevant references for inclusion. Studies that were not yet included in the study and that met the inclusion criteria were also selected. This was done to satisfy the cumulative character of meta-analytic studies on the same topic, which meta-analyses often lack (Bennett, 2011).

4.2.2 Inclusion Criteria

In order for studies to be included in the current meta-analysis, they had to meet the following criteria: 1) the study was published in a journal article, scientific book, book section, or dissertation; 2) the study was published in the English language; and 3) the study compared the effects of different types of (standardized, response-based) feedback in a computer-based assessment or computer-based learning environment on the learning outcomes of individual students in terms of achievement measured in a quantitative way.

Criterion 1 was used to include all the high-quality publications. Conference proceedings, (technical) reports, and meeting abstracts were excluded in this selection step. These types of documents were omitted because it was unclear whether they had been subjected to peer review, which is a generally accepted criterion for ensuring scientific quality. Unpublished documents, such as master theses, were also not included in this meta-analysis. Criterion 2 was set to obtain studies that had a high degree of accessibility. Criterion 3 was used to select only studies relevant for the purpose of this meta-analysis. In this selection step, studies that did not examine the effects of feedback in a computer environment were excluded. In addition, studies were excluded that did not use a comparison group (that received another type of feedback or no feedback) or did not measure the students' learning outcomes in a quantitative way. Furthermore, the feedback had to be response-based, meaning that the computer provided students with feedback based on their response to an item. Thus, the feedback had to be item-specific and identical for all students in the experimental group based on the input response. Studies that used feedback that was generated by a human being and that therefore differed for students within one feedback condition were excluded. No restrictions were made with regard to the subject matter, level of education, or type of computer-based environment.

Experimental and control groups had to contain at least 10 participants each. We did not make very strict restrictions with regard to the nature of the control groups. They could have received no feedback, KR only, or KCR only. Furthermore, the study had to report sufficient quantitative information in order to make it possible to estimate the effect size statistic. Whenever a small amount of essential information was missing, the authors were contacted through e-mail.

Two researchers conducted selection steps one and two independently. After each selection step, the authors compared their judgements and discussed and resolved any differences. The overall agreement rate was 92.85% for selection step one and 98.84% for selection step two. In selection step three, 58% of the studies were judged on relevance for the selection criteria by two researchers; the other 42% was judged by one researcher. The overall agreement rate was 92.19%.

The full text versions of studies that met the inclusion criteria were retrieved. Studies that were not available in their full text versions through the library facilities at hand, including interlibrary loans, were requested from the author. Studies that could not be retrieved in their full text versions were not included in the meta-analysis.

4.2.3 Coding Procedures

The coding form (Appendix 4A) was based on a method proposed by Lipsey and Wilson (2001) and by thoroughly scanning the literature on this topic. The form was evaluated multiple times in repetitive cycles in consultation with the coders and specialists on this topic. The form contains multiple sections, such as study descriptors, sample descriptors, methods and procedures (including study quality), and effect size information.

Coding was conducted by four researchers independently. The first author coded all studies, and 40% of the studies and 31% of all effect sizes were double-coded by one different coder. The agreement rate between the coders was .81 for all studies and .81 for all effect sizes. In some studies, multiple experimental conditions existed in which participants received the same type of feedback under different conditions. For example, a study compared the effects of KCR under program-control conditions and learner-control conditions (Corbalan, Kester, & Van Merriënboer, 2009). In this study, mean scores of groups that received the same type of feedback under different conditions that were not relevant to the present meta-analysis were combined, using their weighted means and pooled standard deviations.

4.2.4 Statistical Methods

Meta-analysis is a method that enables the combination and summary of quantitative information from different studies focusing on the same research question. Using information from several studies gives the opportunity to provide a structured summary of a specific research topic and to find relationships between variables that otherwise would not be detected. A central step in meta-analysis is to make the data from each separate study comparable. This can be done by using an effect size measure. Effect sizes express the effectiveness of the variable of interest found in each study between a treatment and control group in terms of standard deviation units (Lipsey & Wilson, 2001). The experimental and control groups in each study always took the same post-test. Usually, estimates of effect sizes obtain a plus sign if the treatment is better than the control group and vice versa (Rosenthal, 1994). We used the standardized mean differences (see Equation 1) effect size statistic.

$$ES = \frac{\overline{X}_{G1} - \overline{X}_{G2}}{s_{pooled}} \qquad (1)$$

This statistic uses the contrast between the treatment and control group divided by the pooled standard deviation (see Equation 2).

$$s_{pooled} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}} \qquad (2)$$

Hedges (1981) showed that effect sizes calculated for studies with small samples might be biased. An unbiased alternative estimator developed by Hedges (1981) (see Equation 3) was used to correct for this bias.

$$ES' = \left[1 - \frac{3}{4(n_1 + n_2 - 9)}\right] ES \qquad (3)$$

This estimator has the following standard error:

$$SE = \sqrt{\frac{n_{G1} + n_{G2}}{n_{G1}n_{G2}} + \frac{(ES')^2}{2(n_{G1} + n_{G2})}} \qquad (4)$$

The adjusted effect sizes have been checked for potential outliers by evaluating whether certain effect sizes differed more than three *SD*s from several group mean effect sizes, as they might have too large an impact on the summary statistics or even create a bias.

In meta-analysis, one might assume that all included studies estimate the same population mean. Models under this assumption are called *fixed effects models*. In this case, differences between effect size estimates and the true population mean can only be attributed to sampling error (Shadish & Haddock, 1994). The homogeneity test assesses if the observed variability in the results are more heterogeneous than expected from sampling variance alone. This test is based on the Q statistic (Lipsey & Wilson, 2001), which is distributed as a chi-square distribution with the number of effects sizes minus one degree of freedom (Hedges & Olkin, 1985). If the homogeneity of study results is rejected, *random* and *mixed effects* models are alternatives for the fixed effects model. In the former, studies are assumed to estimate a distribution of effect sizes, whereas in the latter, variance in effect sizes can be seen as both having a random and a systematic component. Study characteristics can be used to account for the systematic part of the differences between effects sizes estimates.

Because the included studies have different sample sizes, the estimates from these studies also differ in precision. A common strategy in meta-analysis is therefore to attach different weights to studies depending on the level of precision of the studies. Optimal weights are constructed using the standard error of the effect size (Hedges, 1982; Hedges & Olkin, 1985) (see Equation 5).

$$W = \frac{1}{SE^2} \qquad (5)$$

These weights are used in the fixed effects models. In random effects models, these weights are supplemented by a constant that represents the variability across population effects (see Equation 6).

$$W = \frac{1}{SE^2 + v_\theta} \qquad (6)$$

It might be that studies containing non-significant results about feedback are less easily published than studies with significant results. If that is the case, this meta-analysis would not cover the complete population of studies and would be potentially biased. This issue was addressed with the *fail-safe N* method (Orwin, 1983), which estimates how many studies reporting non-significant effects sizes would be needed in order to get a specific result combined with the studies in this meta-analysis.

In order to assess if effects sizes of studies with different characteristics differ in their mean effect sizes, we used analysis of variance for effect sizes. To establish the relationships between effect sizes and several study characteristics simultaneously, we performed a weighted regression analysis. All analyses were carried out using adjusted SPSS macros written by Wilson (version 2005.05.23).

4.3 Results

4.3.1 Results of the Selection Process

In the primary literature search conducted in May 2011, 2,318 publications were retrieved. After removing duplicate publications, 1,609 unique publications remained. These publications were subsequently subjected to the selection criteria. The results of the selection process can be found in Table 4.1.

Table 4.1
Results of the Selection Process in the Primary Search

Selection step	Total before selection step	Total excluded	Total after selection step
1	1,609	405	1,204
2	1,204	24	1,181
3	1,182	1,057	125
Final	125	89	36

In total, 125 studies retrieved in the primary search appeared to match the selection criteria based on their titles and abstracts. However, some of the selected studies (*n* = 86) were excluded after retrieving their full text versions because they did not report sufficient information, did not contain a sufficient sample size, or did not compare different types of (standardized, response-based) feedback. Other studies were excluded because their full text versions could not be obtained (*n* = 3). For two studies, both a dissertation and journal article were included in the selection. For these cases, the dissertations were excluded from the selection. The primary search resulted in 36 relevant studies for this meta-analysis.

Using the ancestry approach (White, 1994), another 18 studies were retrieved, and 4 met the inclusion criteria. Six of the studies that had not yet been retrieved in the primary and secondary searches and were included in an existing review study appeared to be relevant. However, none of these studies were eligible for inclusion. Thus, the secondary and tertiary searches resulted in another four eligible studies, which produced a total of 40 studies for this meta-analysis. From these 40 studies, 70 effect sizes were obtained. The number of effect sizes will be indicated by *k* in the remainder of this article.

4.3.2 Characteristics of the Selected Studies

The studies in this meta-analysis were published between 1968 and 2012. Of the 40 studies, 30 appeared as a journal article and 10 were published in the form of a doctoral dissertation. The majority of studies were conducted at a university, college, or in adult education (*n* = 32, *k* = 56). The number of studies that have been carried out in secondary education was six (*k* = 12), and two studies were conducted in primary education (*k* = 2). The studies were categorised based on their subject or content area. Four categories were distinguished: 1) social sciences (*n* = 5, *k* = 9), e.g., psychology and education, 2) mathematics (*n* = 6, *k* = 8), e.g., introductory statistics and algebra, 3) science, biology, and geography (*n* = 18, *k* = 29), e.g., chemistry, climate change, and medical education, and 4) languages (*n* = 13, *k* = 24), e.g., vocabulary and Spanish as a second language. The majority of the effect sizes have been derived from studies that assigned the subjects randomly to the experimental or control condition (*k* = 52). The other effect sizes come from studies that assigned their subjects randomly by class (*k* = 4) or by matching (*k* = 12), used a non-random assignment procedure (*k* = 1), or used some other procedure for the assignment of subjects (*k* = 1).

4.3.3 Effect Sizes

The sample size in the primary studies ranged from 24 to 463, with an average sample size of 106.66 (*SD* = 102.83). Given the relatively small sample size of some studies, a correction for an upwards bias was applied, which resulted in a corrected effect size noted as ES' (Lipsey & Wilson, 2001). This meta-analysis contained 70 effect sizes obtained from 40 unique studies. The effect sizes ranged from -0.78 to 2.29. Table 4.2 shows the unweighted effect sizes ordered from smallest to largest per feedback type.

Table 4.2
Author(s) and Publication Year, Feedback Type, Feedback Timing, Sample Size, Level of Learning Outcomes, and Unweighted Effect Sizes

Author(s) and publication year	Feedback type	Feedback timing	Sample size	Level(s) of learning (L / H)*	Unweighted ES'
Neri, Cucchiarini, & Strik (2008)	KR	Immediate	25	L	-0.78
Morrison, Ross, Gopalakrishnan, & Casey (1995)	KR	Immediate	97	L	-0.10
Rosa & Leow (2004)	KR	Immediate	50	H	-0.01
Epstein (1997)	KR	Immediate	65	L	-0.01
Vispoel (1998)	KR	Immediate	293	L	0.09
Rosa & Leow (2004)	KR	Immediate	50	L	0.24
Cameron & Dwyer (2005)	KR	Delayed	150	L	0.28
Morrison, Ross, Gopalakrishnan, & Casey (1995)	KR	Immediate	98	L	0.44
Pridemore & Klein (1995)	KCR	Immediate	138	L	-0.52
Ifenthaler (2010)	KCR	Delayed	74	L	-0.27
Clariana & Lee (2001)	KCR	Immediate	108	L	-0.07
Lin (2006)	KCR	Immediate	388	L + H	0.10
Roos, Wise, & Plake (1997)	KCR	Immediate	363	H	0.17
Morrison, Ross, Gopalakrishnan, & Casey (1995)	KCR	Immediate	99	L	0.24
Corbalan, Kester, & Van Merriënboer (2009)	KCR	Immediate	118	H	0.26
Papa, Aldrich, & Schumacker (1999)	KCR	Immediate	108	L + H	1.03
Valdez (2009)	KCR	Immediate	84	L	2.28
Hall, Adams, & Tardibuono (1968)	EF	Immediate	24	L	-0.34
Murphy (2007)	EF	Delayed	101	L	-0.29
Moreno (2007)	EF	Immediate	59	H	-0.26
Moreno (2007)	EF	Immediate	59	H	-0.23
Munyofu (2008)	EF	Immediate	121	L	-0.21
Sanz & Morgan-Short (2004)	EF	Immediate	33	H	-0.10
Munyofu (2008)	EF	Delayed	120	L	-0.06
Huang (2008)	EF	Immediate	50	H	0.00
Valdez (2009)	EF	Immediate	84	L	0.03
Moreno (2007)	EF	Immediate	59	H	0.05
Xu (2009)	EF	Immediate	43	L	0.07
Gordijn & Nijhof (2002)	EF	Immediate	452	L	0.08
Gordijn & Nijhof (2002)	EF	Immediate	434	L	0.09
Wager (1983)	EF	Delayed	42	L	0.10
Merril (1987)	EF	Immediate	154	L	0.10

Wager (1983)	EF	Delayed	39	L	0.11
Moreno (2004)	EF	Immediate	55	L	0.12
Sanz & Morgan-Short (2004)	EF	Immediate	33	L	0.13
Murphy (2010)	EF	Delayed	267	L	0.17
Mazingo (2006)	EF	Immediate	75	L	0.18
Pridemore & Klein (1995)	EF	Immediate	138	L	0.24
Gordijn & Nijhof (2002)	EF	Immediate	84	L	0.25
Nagata (1993)	EF	Immediate	34	H	0.27
Cameron & Dwyer (2005)	EF	Delayed	148	L	0.29
Mazingo (2006)	EF	Immediate	74	L	0.32
Moreno & Valdez (2005)	EF	Delayed	31	L	0.33
Kopp, Stark, & Fischer (2008)	EF	Immediate	153	L + H	0.35
Epstein (1997)	EF	Immediate	65	L	0.41
Epstein (1997)	EF	Immediate	64	L	0.42
Collins, Carnine, & Gersten (1987)	EF	Immediate	28	L	0.50
Cameron & Dwyer (2005)	EF	Delayed	150	L	0.57
Lee, Lim, & Grabowski (2010)	EF	Immediate	148	L + H	0.62
Collins, Carnine, & Gersten (1987)	EF	Immediate	28	H	0.62
Sanz & Morgan-Short (2004)	EF	Immediate	33	H	0.64
Corbalan, Paas, & Cuypers (2010)	EF	Immediate	34	L + H	0.68
Moreno & Valdez (2005)	EF	Delayed	31	H	0.69
Lipnevich & Smith (2009)	EF	Delayed	463	L + H	0.75
Pridemore & Klein (1995)	EF	Immediate	138	L	0.76
Nagata & Swisher (1995)	EF	Immediate	32	H	0.76
Rosa & Leow (2004)	EF	Immediate	49	H	0.80
Rosa & Leow (2004)	EF	Immediate	67	H	0.83
Moreno (2004)	EF	Immediate	49	L	0.85
Rosa & Leow (2004)	EF	Immediate	67	L	0.85
Bowles (2005)	EF	Immediate	54	H	0.93
Narciss & Huth (2006)	EF	Immediate	50	H	0.97
Rosa & Leow (2004)	EF	Immediate	49	L	1.10
Moreno (2004)	EF	Immediate	49	H	1.18
Kim & Phillips (1991)	EF	Immediate	24	L	1.23
Kramarski & Zeichner (2001)	EF	Immediate	186	H	1.28
Moreno (2004)	EF	Immediate	55	H	1.35
Pridemore & Klein (1991)	EF	Immediate	47	L	1.59
Valdez (2009)	EF	Immediate	84	L	2.25
Lee, Lim, & Grabowski (2010)	EF	Immediate	148	L + H	2.29

Note. L = lower-order, H = higher-order.

It can be concluded from Table 4.2 that the majority of the effect sizes concern the effects on lower-order learning outcomes (k = 43). Other effect sizes (k = 27) indicated the effects on higher-order learning outcomes or a combination of lower-order and higher-order learning outcomes. Furthermore, the results showed that the majority of the effect sizes were concerned with immediate feedback (k = 58), and only a small number of the effect sizes were associated with feedback that was delivered with a delay (k = 12).

In the majority of cases (k = 61), students only received feedback at one assessment occasion. In two cases, students received feedback on their assessment results twice. In the remainder of the cases (k = 7), students were given feedback three times or more.

Table 4.2 shows that the number of effect sizes that were negative was 12, and for each feedback type, there were both negative and positive effect sizes. The distribution of the effect sizes and their 90% CI is shown in Figure 4.1. No cases were found that deviated more than three SDs from several group averages. Therefore, we decided not to exclude or adjust any effect size.

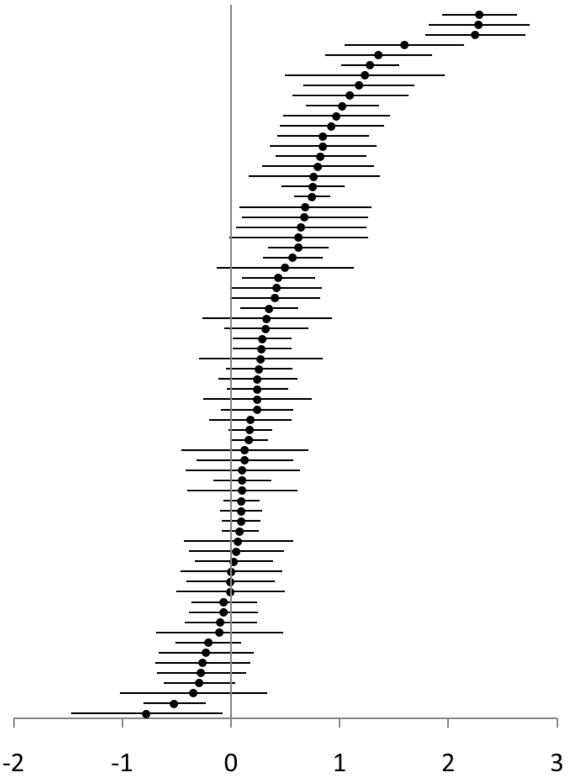


Figure 4.1. The distribution of the unweighted effect sizes and their 90% CI.

The analyses for homogeneity showed that the observed variation in the effects was more heterogeneous than what might be expected from sampling variance alone (χ^2= 446, *df* = 69, $p < 0.001$). This result suggested that a fixed model may not be suitable. Therefore, weights were supplemented by a random component (see Equation 6). No indication was found for publication bias. In the case of this meta-analysis one would need 76 extra studies with an overall effect size of 0 to obtain a small mean weighted effect size of 0.2 (Hattie, 2009). Furthermore, the ratio of the number of included effect sizes and number of studies (70/40) was considered too small to perform multilevel analyses in order to account for the non-independence of effect sizes within studies. For this reason, the analyses were performed using regular linear regression.

We computed the mean weighted effect sizes for various relevant variables. The results are shown in Table 4.3. The mean weighted effect size in Table 4.3 was the associated mean effect size for each category, followed by the corresponding 90% CI and *p*-value. Table 4.3 shows that of the three feedback types, the effects for KR were smallest and those for EF were largest. With regard to feedback timing, the effects for immediate feedback were larger than for delayed feedback. Furthermore, the effect sizes appeared to be larger for higher-order learning outcomes than for lower-order learning outcomes. The effects found in university or college settings were somewhat larger than those found in primary and high school settings. The effects also seemed to differ with respect to the various subject areas from which they were derived. The effects of studies in mathematics were very large, whereas the effects of studies in social sciences and science were medium, and those in languages were small.

Table 4.3
ANOVA Results

Variable	Category	*k*	-90% CI	Mean weighted ES'	+90% CI	*p*
Feedback type	KR	8	-0.28	0.05	0.39	.79
	KCR	9	0.02	0.33	0.63	.07
	EF	53	0.36	0.49	0.62	.00
Feedback timing	Immediate	58	0.33	0.46	0.59	.00
	Delayed	12	-0.05	0.22	0.49	.19
Level of learning outcomes	Lower	43	0.17	0.31	0.45	.00
	Higher & lower + higher	27	0.41	0.59	0.78	.00
Education level	University/College	56	0.31	0.44	0.57	.00
	Primary + high school	14	0.08	0.34	0.59	.03
Subject	Social sciences	9	0.15	0.46	0.77	.02
	Mathematics	8	0.61	0.93	1.26	.00
	Science	29	0.23	0.40	0.57	.00
	Languages	24	0.05	0.25	0.44	.03

Next, the effect sizes per feedback type were computed for each of the control groups (see Table 4.4). The theoretical framework led us to assign the conditions in the primary studies to the experimental and control conditions for computing the effect sizes. In this study, we chose to compare feedback types to the possible other feedback conditions with which they might be contrasted. However, Table 4.4 shows that there was considerable variation in the magnitude of the effect sizes with respect to the different control groups. Nevertheless, given the small sample size within this meta-analysis, it was deemed not meaningful to run separate analysis for each possible control group.

Table 4.4
Effect Sizes per Feedback Type per Control Condition

Feedback type	Control condition	*k*	-90% CI	Mean weighted ES'	+90% CI
KR	No feedback	8	-0.28	0.05	0.39
	KR	-	-	-	-
	KCR	-	-	-	-
KCR	No feedback	7	0.04	0.48	0.92
	KR	2	-1.00	-0.17	0.66
	KCR	-	-	-	-
EF	No feedback	9	0.31	0.61	0.91
	KR	21	0.33	0.54	0.74
	KCR	23	0.20	0.39	0.60

It was expected that KR and KCR would have a small to moderately positive effect (between 0.2 and 0.6) on lower-order learning outcomes (Hypothesis 1). Furthermore, we predicted that KR and KCR would have virtually no effect (below 0.2) on higher-order learning outcomes (Hypothesis 2). With respect to EF, it was expected that it would have a moderate to large positive effect (at least 0.4) on lower-order learning outcomes (Hypothesis 3) and also a moderate to large effect on higher-order learning outcomes (Hypothesis 4). Table 4.5 shows the mean weighted effect sizes and their 90% CI of KR, KCR, and EF by level of learning outcomes.

Table 4.5

Mean weighted Effect Sizes of KR, KCR, and EF by Level of Learning Outcomes

Feedback type	Level of learning outcomes	*k*	-90% CI	Mean weighted ES’	+90% CI
KR	L	7	0.00	0.12	0.24
	H and L + H	1	-0.51	-0.01	0.49
KCR	L	5	-0.25	0.31	0.87
	H and L + H	4	-0.23	0.38	0.99
EF	L	31	0.21	0.37	0.53
	H and L + H	22	0.47	0.67	0.87

Note. L = lower-order, H = higher-order, L + H = lower-order and higher-order.

Table 4.5 also indicates that the mean weighted effect sizes for KR on lower-order learning outcomes was 0.12, which is somewhat lower than the expected effect of at least 0.2. The effects of KCR on lower-order learning outcomes are in line with the expectations, namely ES’ = 0.31. The 90% CIs of both KR and KCR fall within the range 0.2–0.6. However, the intervals were large due to the small number of observations. These large intervals indicate that we cannot meaningfully test Hypothesis 1. The effects on higher-order learning outcomes of KR were as expected; however, the effects of KCR were larger than expected. Again, the 90% CIs were large, and we had insufficient power to meaningfully test Hypothesis 2.

Hypothesis 3 was not rejected because the effect of EF on lower-order learning outcomes was 0.37, and the upper limit of the CI was above 0.4. The effects of EF on higher-order learning outcomes were somewhat larger than 0.4, meaning that Hypothesis 4 was not rejected.

The effects of EF seemed promising, but the nature of EF varies widely. Therefore, we categorised feedback effects based on the level (Hattie & Timperley, 2007) at which they were intended to gain more insight into which method for providing EF was most effective (see Table 4.6).

Table 4.6

The Mean weighted Effect Sizes of EF for each Feedback Level

Feedback level	*k*	-90% CI	Mean weighted ES'	+90% CI
Task	4	-0.46	-0.06	0.34
Regulation	1	-0.65	0.08	0.81
Self, task, and process	2	-0.31	0.25	0.82
Task and process	41	0.36	0.50	0.63
Task and regulation	4	0.63	1.05	1.47
Task, process, and regulation	1	0.52	1.29	2.04

The results showed that the majority of the effect sizes (*k* = 41) involved the task and process level, which showed moderately large effects (ES' = 0.50). The largest effects were found for EF at the task and regulation level (ES' = 1.05, *k* = 4), and at the task, process, and regulation level (ES' = 1.29, *k* = 1), although the number of effects was small.

We expected that there would be an interaction effect between feedback timing and the level of learning outcomes (Hypothesis 5). Figure 4.2 shows that the directionality of the effects was consistent with Hypothesis 5, with immediate feedback being more effective for lower-order learning outcomes and vice versa.

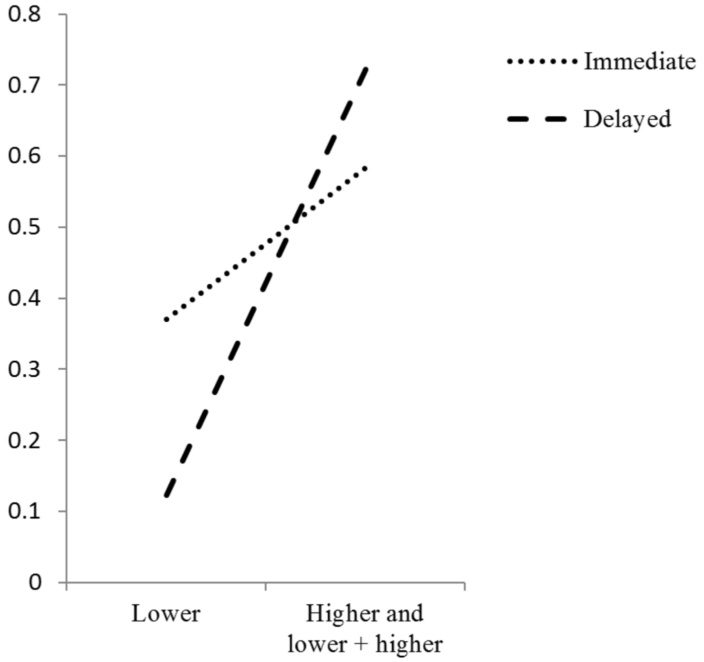


Figure 4.2. Effects of immediate and delayed feedback by level of learning outcomes.

There appeared to be no significant interaction effect between the level of learning outcomes and feedback timing (*z* = 0.82, *p* = 0.41). Therefore, Hypothesis 5 was rejected. The lack of power in the analysis might be a reason for the lack of statistical significance. For example, the combination of higher-order learning outcomes and delayed feedback contained only two observations.

To evaluate the relationships between study attributes and effects sizes simultaneously, a weighted regression analysis was conducted. Table 4.7 summarizes the results of this analysis.

Table 4.7
Weighted Regression Results, random Intercept, fixed Slopes Model

Dummy coded variables	-90% CI	*B*	+90% CI	*SE*	*p*
Constant	-0.38	-.06	0.26	0.19	.762
EF	0.05	.40	0.75	0.21	.057
KCR	-0.40	.03	0.46	0.26	.904
Higher-order learning outcomes	-0.11	.12	0.35	0.14	.382
Delayed timing	-0.65	-.35	-0.04	0.18	.059
Primary + high school	-0.65	-.35	-0.05	0.18	.051
Social sciences	0.06	.41	0.76	0.21	.054
Mathematics	0.33	.70	1.06	0.22	.002
Science	0.03	.30	0.56	0.16	.065

The regression model explained 25% of the total variance across effect size estimates. Table 4.7 shows that especially the feedback type EF and the subject areas of social sciences, mathematics, and science had a positive impact on the effect size estimates. The effects of mathematics on the effect size estimates were strikingly large. Delayed feedback and primary + high school had a negative effect on the estimates.

4.4 Discussion

The purpose of this meta-analysis was to gain insight into the effects of various methods for providing feedback to students in a computer-based environment in terms of students' learning outcomes. The major independent variable in this meta-analysis was feedback type (Shute, 2008). Furthermore, the effects of various moderator variables that seemed relevant given the literature on feedback effects, such as timing (Shute, 2008) and level of learning outcomes (Van der Kleij et al., 2011), were investigated.

The 70 effect sizes in this meta-analysis were derived from 40 studies and expressed the difference in the effects of one feedback type compared to another feedback type or no feedback at all in terms of a test score. The effect sizes ranged from -0.78 to 2.29. Because of the heterogeneous nature of the collection of effect sizes, a mixed model was used in the analyses. The majority of the effect sizes (k = 53) concerned the investigation effects of EF in contrast to KCR, KR, or no feedback. The results suggested that EF was more effective than KR and KCR. The mean weighted effect size for EF was 0.49, which can be considered as a moderately large effect. The mean weighted effect size for KCR (k = 9) was 0.32, which is considered small to moderate. The effect size for KR (k = 8) was very small at 0.05.

KR and KCR were expected to have a small to moderately positive effect (between 0.2 and 0.6) on lower-order learning outcomes (Hypothesis 1). Due to the small number of observations, we could not meaningfully test Hypothesis 1, but the limited information available did not contradict our expectations. In addition, we expected that KR and KCR would have virtually no effect (below 0.2) on higher-order learning outcomes (Hypothesis 2). The effects of KCR were slightly larger than expected: 0.38. However, we had insufficient power to reject Hypothesis 2. EF was expected to have a moderate to large positive effect (at least 0.4) on both lower-order learning outcomes (Hypothesis 3) and higher-order learning outcomes (Hypothesis 4). Hypothesis 3 and Hypothesis 4 were not rejected, and the effects of EF on higher-order learning outcomes (ES'= 0.67) appeared to be larger than the effects on lower-order learning outcomes (ES'= 0.37).

The effects of EF seemed promising, although the nature of EF varies widely. By categorising feedback effects based on the level at which they are aimed, we attempted to gain more insight into which method for providing EF was most effective. However, the majority of the EF was aimed at the task and process level (k = 41), which makes it difficult to draw any generalizable conclusions regarding the effects of the different feedback levels. The mean weighted effect size of EF at the task and process level was 0.50. In this meta-analysis, the effect sizes from EF at the task level (k = 4) were lowest (ES'= -0.06). The effects of EF at the task and regulation level (k = 4, ES'= 1.05) and at the task, process, and regulation level (k = 1, ES'= 1.29) were highest. These effects can be regarded as very large, which suggests that more research is warranted regarding the effects of feedback at the task and/or process level in combination with the regulation level. The results of this meta-analysis are in line with the results of the systematic review by Van der Kleij et al. (2011), which suggested that the effects of EF at the regulation level are promising but have not been researched to a great extent. Consistent with the literature on feedback effects (e.g., Hattie & Timperley, 2007), adding feedback that is not task related but is aimed instead at the characteristics of the learner seemed to impede the positive effects of EF (k = 2, ES'= 0.25). It must be mentioned, however, that the number of studies examining feedback that is aimed at the self-level in a computer-based environment is fortunately low.

Furthermore, it was hypothesized that there would be an interaction effect between feedback timing and the level of learning outcomes (Hypothesis 5). Hypothesis 5 was rejected because there appeared to be no significant interaction effect between the level of learning outcomes and feedback timing. The directionality of the effects was, however, consistent with Hypothesis 5. Namely, it was expected that immediate feedback would be more effective for lower-order learning outcomes and vice versa (Shute, 2008). However, possibly due to a lack of power, statistical significance was not reached. More research is needed to shed light on the possible interaction between feedback timing and the level of learning outcomes.

In addition, to evaluate the relationships between study attributes and effect sizes simultaneously, a weighted regression analysis was conducted. This analysis pointed out that delayed feedback and primary and high school negatively affected the ES' estimates. Furthermore, EF and the subject areas social sciences, science, and especially mathematics positively affected the ES' estimates. Moreover, the effect of mathematics was strikingly high. However, this effect was only based on eight studies. Of these studies, only two did not include EF, which makes it likely that the high effects are undeservedly attributed to the subject mathematics. Furthermore, it must be mentioned that the literature does not show any consistent positive effects of feedback in mathematics (e.g., Bangert-Drowns et al., 1991;

Kingston & Nash, 2011). Therefore, the positive results of studies conducted in the field of mathematics must be interpreted with caution.

A limitation of this meta-analysis—and of review studies in general—is the impossibility of retrieving all relevant studies. Moreover, this meta-analysis only included published work with the exception of unpublished doctoral dissertations. The authors deliberately chose to exclude unpublished sources because there is no way of objectively retrieving these works. Besides, it was also unclear whether they had been subjected to peer review, which is a generally accepted criterion for ensuring scientific quality. With that exception, no strict requirements were established with regard to the quality of the included studies. Studies with low quality were usually judged as being of low quality as a result of their limited sample size, which automatically would result in a low weight compared to studies of a higher quality. Furthermore, many studies had to be excluded because they did not provide sufficient information for computing an effect size.

For some studies, multiple effect sizes were coded because the studies included multiple experimental groups. Therefore, the effect sizes within the dataset were not completely independent. Multilevel analysis can address this nested data structure appropriately, but in our case, the ratio of effect sizes within studies (70/40) was too small to conduct a multilevel analysis.

Another limitation of this meta-analysis was that it included insufficient data to meaningfully compare feedback effects across school types. Namely, the majority of the studies was conducted at universities, colleges, or other places of adult education. Given the low number of studies in secondary education (n = 6) and the even lower number of studies in primary education (n = 2), the degree to which the conclusions of this meta-analysis apply to young learners is questionable. The results suggest that there is reason to believe that feedback mechanisms function differently within these various school types. Moreover, providing feedback in the form of text may not be appropriate for younger learners since their reading abilities might not be sufficiently developed to fully understand the feedback and subsequently use it. However, current technology makes it possible to provide feedback in many ways (Narciss & Huth, 2006). For example, the feedback could be channelled to the students by audio, graphical representations, video, or in a game. Unfortunately, only a limited number of studies included in this meta-analysis used multimedia feedback (n = 7, e.g., Narciss & Huth, 2006; Xu, 2009), which means that no meaningful comparison could be made in this meta-analysis with respect to feedback mode. The results do suggest, however, that the area of multimedia feedback is one that needs further exploration.

It is striking that in most of the studies in this meta-analysis, the researchers assumed that the learners paid attention to the feedback provided. The results of recent research suggest, however, that in a computer-based environment, some students tend to ignore written feedback (e.g., Timmers & Veldkamp, 2011; Van der Kleij et al., 2012). Variables like motivation and learners' perceived need to receive feedback play an important role in how feedback is received and processed (Stobart, 2008). These variables therefore intervene with other variables that contribute to feedback effectiveness, such as type and timing. Nevertheless, based on currently available research, it is not possible to examine the interplay of these variables thoroughly. In the experiments by Timmers and Veldkamp and Van der Kleij et al., the time students chose to display feedback for each item was logged as an operationalisation of time spent examining the feedback. Using eye-tracking technologies could reveal more detailed images of student behaviour in examining feedback in computer-based environments.

Furthermore, the effects reported in this meta-analysis were all short-term effects, measured using a post-test that was administered either immediately or shortly after the feedback intervention. For feedback to realise its full formative potential, however, continuous feedback loops are needed throughout the entire learning process.

The literature on the effectiveness of feedback suggests that the complex relationships between the feedback intervention, the task, the learning context, and the characteristics of the learner impact the magnitude of feedback effects (Shute, 2008). However, the primary studies that have been published to date have reported insufficient data to meaningfully examine these complex relationships. For example, research suggests that the initial ability levels of the learner affect feedback effectiveness (e.g., Hattie & Gan, 2011; Smits, Boon, Sluijsmans, & Van Gog, 2008), but only a small number of studies in this meta-analysis reported information about the initial ability of the learners. Hattie and Gan have suggested that feedback needs to be appropriate for the level at which the learner is functioning. In a computer-based environment, the mechanisms of computerised adaptive testing (CAT; Wainer, 2000) are very promising in this respect. If not only the item selection but also the selection of the specific feedback to the item response could be selected based on the current ability of the learner, feedback can perhaps fulfil its formative potential to a greater extent (Veldkamp, Matteucci, & Eggen, 2011). In addition, providing feedback that is gradually becoming more elaborated in an interactive manner could be used to adapt the feedback to the needs of learners. This kind of feedback is called intelligent tutoring feedback (ITF; Narciss, 2008). Digital learning environments could become even more powerful when the feedback includes a diagnostic component, meaning that it has been designed to address, prevent, or correct misconceptions or frequently made errors. However, research is needed to explore the potentials of computer-based adaptive learning environments and adaptive feedback mechanisms.

Moreover, the degree in which the psychometrical properties of the instruments used have been reported in the primary studies is strikingly low. For example, only a few studies reported the reliability (α) of the post-test. Furthermore, the number of items used in the post-test was also not reported in all studies, and the studies that did report the number of items used a strikingly low number. These assessments are insufficiently reliable for making well-grounded claims about the differences in the effects of various feedback types. Also, only a few studies reported the difficulty level of the items in the assessments. In future research, it is recommended that these psychometrical properties be reported because they can be used to more accurately weight the effect sizes. These data could also be used to examine the relationships between optimal item difficulty in formative computer-based assessments and the initial ability of learners since difficult items leave room for more opportunity to learn from feedback than easier items (Van der Kleij et al., 2011). Ultimately, these insights could be used to optimise computer adaptive learning environments (e.g., Wauters, Desmet, & Van den Noortgate, 2010). Finally, in future research, it is recommended that larger groups of participants be used.

The results of this meta-analysis consistently showed that more elaborated feedback led to higher learning outcomes than simple feedback. This finding has important implications for designers of educational software/computer-based learning environments. Furthermore, the insights gained from this meta-analysis could help educational practitioners to make well-informed choices with respect to digital learning tools.

From this study, various areas that need further attention in the literature on feedback in computer-based environments can be identified. This meta-analysis made an initial attempt to gain insight into the effectiveness of various methods for providing EF. Nonetheless, more research is needed that takes into account the specific characteristics of EF interventions. In future experiments, it is advised that larger sample sizes be used and that the psychometrical properties of the instruments be reported. Future investigations should account for and report on characteristics of the feedback, the task, the learning context, and the characteristics of the learners in order to move the research field further. Using eye tracking could also provide usable insight into how students react to the feedback provided, which might differ based on the correctness of their answer to the item. The results of this meta-analysis clearly showed that there is a need for research on this topic in primary education settings. In addition, the authors recommend that the value of multimedia feedback be investigated in future research.

References

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

Assessment Reform Group (1999). *Assessment for learning: Beyond the black box.* Retrieved from http://assessmentreformgroup.files.wordpress.com/2012/01/beyond_blackbox.pdf

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research, 13*, 111–127. doi:10.2190/9LMD-3U28-3A0G-FTQT

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213–238. doi:10.3102/00346543061002213

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

Black, P., & Wiliam, D. (1998a). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–48.

Black, P., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. London, UK with: King's College.

Black, P., & Wiliam, D. (1998c). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*, 7–74. doi:10.1080/0969595980050102

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher 13*(6), 4–16.

* Bowles, M. A. (2005). *Effects of verbalization condition and type of feedback on L2 development in a CALL task* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3175847)

Briggs, D. C., Ruiz-Primo, M. A., Furtak, E., Shepard, L., & Yin, Y. (2012). Meta-analytic methodology and inferences about the efficacy of formative assessment. *Educational Measurement: Issues and Practice*, *31*, 13–17. doi:10.1111/j.1745-3992.2012.00251.x

* Cameron, B., & Dwyer, F. (2005). The effect of online gaming, cognition and feedback type in facilitating delayed achievement of different learning objectives. *Journal of Interactive Learning Research, 16*(3), 243–258.

* Clariana, R. B., & Lee, D. (2001). The effects of recognition and recall study tasks with feedback in a computer-based vocabulary lesson. *Educational Technology Research and Development, 49*, 23–36. doi:10.1007/BF02504913

* Collins, M., Carnine, D., & Gersten, R. (1987). Elaborated corrective feedback and the acquisition of reasoning skills: A study of computer-assisted instruction. *Exceptional Children, 54*(3), 254–262.

* Corbalan, G., Kester, L., & Van Merriënboer, J. J. G. (2009). Dynamic task selection: Effects of feedback and learner control on efficiency and motivation. *Learning and Instruction, 19*, 455–465. doi:10.1016/j.learninstruc.2008.07.002

* Corbalan, G., Paas, F., & Cuypers, H. (2010). Computer-based feedback in linear algebra: Effects on transfer performance and motivation. *Computers and Education, 55*, 692-703. doi:10.1016/j.compedu.2010.03.002

* Epstein, J. I. (1997). *The effects of different types of feedback on learning verbal reasoning in computer-based instruction* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9727464)

* Gordijn, J., & Nijhof, W. J. (2002). Effects of complex feedback on computer-assisted modular instruction. *Computers and Education, 39*, 183–200. doi:10.1016/S0360-1315(02)00025-8

* Hall, K. A., Adams, M., & Tardibuono, J. (1968). Gradient- and full-response feedback in computer-assisted instruction. *Journal of Educational Research, 61*(5), 195–199.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* London, UK: Routledge.

Hattie, J., & Gan, M. (2011). Instruction based on feedback. In P. Alexander & R. E. Mayer (Eds.), *Handbook of research on learning and instruction* (pp. 249–271). New York, NY: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112. doi:10.3102/003465430298487

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of educational statistics, 6*(2), 107–128.

Hedges, L. V. (1982). Estimation of the effect size from a series of independent experiments. *Psychological Bulletin, 92*(2), 490–499.

Hedges, L.V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press.

* Huang, T. O. (2008). *The role of task-specific adapted knowledge of response feedback in algebra problem solving online homework in a college remedial course* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3325192)

* Ifenthaler, D. (2010). Bridging the gap between expert-novice differences: The model-based feedback approach. *Journal of Research on Technology in Education, 43*(2), 103–117.

Jaehnig, W., & Miller, M. L. (2007). Feedback types in programmed instruction: A systematic review. *Psychological Record, 57*(2), 219–232.

* Kim, J. Y. L., & Phillips, T. L. (1991). The effectiveness of two forms of corrective feedback in diabetes education. *Journal of Computer-Based Instruction, 18*(1), 14–18.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*, 28–37. doi:10.1111/j.1745-3992.2011.00220.x

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284. doi:10.1037/0033-2909.119.2.254

* Kopp, V., Stark, R., & Fischer, M. R. (2008). Fostering diagnostic knowledge through computer-supported, case-based worked examples: Effects of erroneous examples and feedback. *Medical Education, 42*, 823–829. doi:10.1111/j.1365-2923.2008.03122.x

* Kramarski, B., & Zeichner, O. (2001). Using technology to enhance mathematical reasoning: Effects of feedback and self-regulation learning. *Educational Media International, 38*, 77–82. doi:10.1080/09523980110041458

* Lee, H. W., Lim, K. Y., & Grabowski, B. L. (2010). Improving self-regulation, learning strategy use, and achievement with metacognitive feedback. *Educational Technology Research and Development, 58*, 629–648. doi:10.1007/s11423-010-9153-6

* Lin, H. (2006). *The effect of questions and feedback used to complement static and animated visualization on tests measuring different educational objectives* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3318901)

* Lipnevich, A. A., & Smith, J. K. (2009). Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied, 15*, 319–333. doi:10.1037/a0017841

Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

* Mazingo, D. E. (2006). *Identifying the relationship between feedback provided in computer-assisted instructional modules, science self-efficacy, and academic achievement* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3263234)

* Merril, J. (1987). Levels of questioning and forms of feedback: Instructional factors in courseware design. *Journal of Computer-Based Instruction*, *14*(1), 18–22.

Miller, T. (2009). *Formative computer-based assessments: The potentials and pitfalls of two formative computer-based assessments used in professional learning programs* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 305048958)

* Moreno, N. (2007). *The effects of type of task and type of feedback on L2 development in call* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3302088)

* Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science, 32*, 99–113. doi:10.1023/B:TRUC.0000021811.66966.1d

* Moreno, R., & Valdez, A. (2005). Cognitive load and learning effects of having students organize pictures and words in multimedia environments: The role of student interactivity and feedback. *Educational Technology Research and Development, 53*, 35–45. doi:10.1007/BF02504796

* Morrison, G. R., Ross, S. M., Gopalakrishnan, M., & Casey, J. (1995). The effects of feedback and incentives on achievement in computer-based instruction. *Contemporary Educational Psychology, 20*, 32–50. doi:10.1006/ceps.1995.1002

Mory, E. H. (2004). Feedback research revisited. In D. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 745–783). Mahwah, NJ: Lawrence Erlbaum Associates.

* Munyofu, M. (2008). *Effects of varied enhancement strategies (chunking, feedback, gaming) in complementing animated instruction in facilitating different types of learning objectives* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3414357)

* Murphy, P. (2007). Reading comprehension exercises online: The effects of feedback, proficiency and interaction. *Language Learning and Technology, 11*(3), 107–129.

* Murphy, P. (2010). Web-based collaborative reading exercises for learners in remote locations: The effects of computer-mediated feedback and interaction via computer-mediated communication. *ReCALL, 22*(2), 112–134.

* Nagata, N. (1993). Intelligent computer feedback for second language instruction. *Modern Language Journal, 77*, 330–339. doi:10.1111/j.1540-4781.1993.tb01980.x

* Nagata, N., & Swisher, M. V. (1995). A study of consciousness-raising by computer: The effect of metalinguistic feedback on second language learning. *Foreign Language Annals, 28*, 337–347. doi:10.1111/j.1944-9720.1995.tb00803.x

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merril, J. J. G. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 125–144). Mahwah, NJ: Lawrence Erlbaum Associates.

* Narciss, S., & Huth, K. (2006). Fostering achievement and motivation with bug-related tutoring feedback in a computer-based training for written subtraction. *Learning and Instruction, 16*, 310–322. doi:10.1016/j.learninstruc.2006.07.003

* Neri, A., Cucchiarini, C., & Strik, H. (2008). The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch. *ReCALL, 20*, 225–243. doi:10.1017/S0958344008000724

Nyquist, J. B. (2003). The benefits of reconstruing feedback as a larger system of formative assessment: A meta-analysis. Unpublished Master's thesis, Vanderbilt University, Nashville, TN.

Orwin, R. G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics, 8*, 157–159. doi:10.3102/10769986008002157

* Papa, F. J., Aldrich, D., & Schumacker, R. E. (1999). The effects of immediate online feedback upon diagnostic performance. *Academic Medicine, 74*(Suppl 10), S16–S18.

* Pridemore, D. R., & Klein, J. D. (1991). Control of feedback in computer-assisted instruction. *Educational Technology Research and Development, 39*, 27–32. doi:10.1007/bf02296569

* Pridemore, D. R., & Klein, J. D. (1995). Control of practice and level of feedback in computer-based instruction. *Contemporary Educational Psychology, 20*, 444–450. doi:10.1006/ceps.1995.1030

Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education, 17*. 1–24. doi:10.1207/s15324818ame1701_1

* Roos, L. L., Wise, S. L., & Plake, B. S. (1997). The role of item feedback in self-adapted testing. *Educational and Psychological Measurement, 57*, 85–98. doi:10.1177/0013164497057001005

* Rosa, E. M., & Leow, R. P. (2004). Computerized task-based exposure, explicitness, type of feedback, and Spanish L2 development. *Modern Language Journal, 88*, 192–216. doi:10.1111/j.0026-7902.2004.00225.x

Rosenthal. R. (1994). Parametric measures of effect size. In Cooper, H. & Hedges, L.V. (Eds.), *The Handbook of Research Synthesis* (pp. 231–244). New York, NY: Russell Sage Foundation.

Sanz, C., & Morgan-Short, K. (2004). Positive evidence versus explicit rule presentation and explicit negative feedback: A computer-assisted study. *Language Learning, 54*, 35–78. doi:10.1111/j.1467-9922.2004.00248.x

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In Cooper, H. & Hedges, L.V. (Eds.), *The Handbook of Research Synthesis* (pp. 261–279). New York, NY: Russell Sage Foundation.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153–189. doi:10.3102/0034654307313795

Smith, P. L., & Ragan, T. J. (2005). *Instructional design* (3rd ed.). New York, NY: Wiley.

Smits, M., Boon, J., Sluijsmans, D. M. A., & Van Gog, T. (2008). Content and timing of feedback in a web-based learning environment: Effects on learning as a function of prior knowledge. *Interactive Learning Environments, 16*, 183–193. doi:10.1080/10494820701365952

Stobart, G. (2008). *Testing times: The uses and abuses of assessment.* Abingdon, UK: Routledge.

Timmers, C. F., & Veldkamp, B. P. (2011). Attention paid to feedback provided by a computer-based assessment for learning on information literacy. *Computers & Education, 56,* 923–930. doi:10.1016/j.compedu.2010.11.007

Thomson Reuters (2010). Thomson Reuters EndNote (Version X4.0). [Computer software]. Available from: http://www.endnote.com/

* Valdez, A. J. (2009). *Encouraging mindful feedback processing: Computer-based instruction in descriptive statistics* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3329482)

Van der Kleij, F. M., Timmers, C. F., & Eggen, T. J. H. M. (2011). The effectiveness of methods for providing written feedback through a computer-based assessment for learning: A systematic review. *CADMO, 19*, 21–39. doi:10.3280/CAD2011-001004

Van der Kleij, F. M., Eggen, T. J. H. M., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers & Education, 58*, 263–272. doi:10.1016/j.compedu.2011.07.020

Veldkamp, B. P., Matteucci, M., & Eggen, T. J. H. M. (2011). Computerized adaptive testing in computer assisted learning? *Communications in Computer and Information Science, 126,* 28–39.

* Vispoel, W. P. (1998). Psychometric characteristics of computer-adaptive and self-adaptive vocabulary tests: The role of answer feedback and test anxiety. *Journal of Educational Measurement, 35*, 155–167. doi:10.1111/j.1745-3984.1998.tb00532.x

* Wager, S.U. (1983). *The effect of immediacy and type of informative feedback on retention in a computer-assisted task* (Unpublished doctoral dissertation). The Florida State University, Tallahassee.

Wainer, H. (Ed.) (2000). *Computerized adaptive testing. A primer* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Wauters, K., Desmet, P., & Van den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: Possibilities and challenges. *Journal of Computer Assisted Learning, 26,* 549–562. doi:10.1111/j.1365-2729.2010.00368.x

White, H. D. (1994). Scientific communication and literature retrieval. In H. Cooper and L. Hedges (Eds.), *The Handbook of research synthesis* (pp. 41–55). NY: Russell Sage Foundation.

Wiliam, D. (2011). What is Assessment for Learning? *Studies in Educational Evaluation*, *37*, 3–14. doi:10.1016/j.stueduc.2011.03.001

* Xu, M. (2009). *An investigation of the effectiveness of intelligent elaborative feedback afforded by pedagogical agents on improving young Chinese language learners' vocabulary acquisition* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3359045)

Appendix 4A. Coding scheme

Varname	Description	Response
IDnr	Study ID number	
NumEs	Number of effect sizes for this study	
Coder	Name coder	1. Fabienne 2. Merijn 3. Marianne 4. Caroline
Cdate	Date of coding (dd-mm-yyyy)	
Publ	Publication form	1. Journal article 2. Book3. Book section 4. Dissertation
Auth	Author(s)	
Year	Year of publication	
Countr	Country in which the study was performed	
Male	Male author	1. Male 0. Female 9. Unknown
Fund	Study sponsorship or funding agency	
Sample descriptors		
Ssize	Sample size (total)	
Mage	Mean age	999 is unknown
SDage	Standard deviation age	999 is unknown
PercF	Percentage female participants	999 is unknown
Edulev	Level of education (descriptive, + grade) Code the level of education and the grade level as specifically as possible. This variable will be grouped later.	
Subj	Subject. Code the subject of the assessments as specifically as possible. This variable will be grouped later.	

Independent variables		
Rq	Research question	
FBtype	Feedback type • KR is only telling whether the answer is correct or incorrect. • KCR is only telling whether the answer is correct or incorrect, and providing the correct answer. • EF is elaborated feedback, everything that provides more information than can KR and KCR	1. KR 2. KCR 3. EF (also include +KR / KCR)
FBlev	Feedback level(s) • task is about the correctness of the answer • process is about the process that has been conducted in order to answer the item • regulation is about processes in the mind of the learner, like self-assessment • self is not related to the task, only at characteristics of the learner, for example praise	1. Task 2. Process 3. Regulation 4. Self 5. Task and Process 6. Task and Regulation 7. Task, process and regulation
FBtime	Feedback timing • immediate is directly after answering an item • delayed is immediately after the assessment • delayed later is everything later	1. Immediate 2. Delayed immediately after assessment 3. Delayed later 9 = Unknown
FBcorr	Was the feedback the same for correctly and incorrectly answered items?	1. Yes 2. No If no, please specify
SN01	Please specify no	
FBway	Way feedback is channelled to students	1. By text 2. By audio 3. By graphics 4. By video 5. In a game 6. In a simulation 7. By an intelligent tutor or animated agent 999 is unknown

Formx	How many times did the students take the formative CBA? In case of a single intervention choose 1.	1. One time 2. Two times 3. Three or more times 4. Students could access the material whenever they wanted 9. Unknown
SpanF	What is the time span in which the formative CBA(s) was / were offered? - The length of the treatment could be related to finding effects. In case of a single intervention choose 1.	1. One day 2. One week 3. Between one and three weeks 4. More than three weeks 5. Unknown
FBchar	Additional feedback characteristics. Try against means they can answer the item again after an initially incorrect answer. This can include one or multiple chances. Overt response means students have to type in the correct answer after receiving feedback. Multiple answers are possible here! These will be grouped/separated later on.	1. No additional characteristics 2. Single try again (KR) 3. Multiple try again (KR) 4. Single try again (hint) 5. Multiple try again (hint) 6. Overt response 7. Worked out solutions 8. Worked out solutions and explanations 9. Explanations 10. Location of error(s) 11. Strategy information 12. Informative tutoring 13. Metacognitive information 14. Demonstration of consequences (visual) 15. Demonstration of correct solution 16. KCR 17. Other (please specify)
S01	Please specify other	
Ittype	Item types in formative CBA. • Close ended items are for example multiple-choice items. • Open ended items are items were students have to construct a response themselves, including fill in the gaps (cloze). • Multiple types means that various item types are combined • Production tasks for example writing tasks.	1. Close ended 2. Open ended (or cloze) 3. Multiple types 4. Production task 9.Unknown

LearnL	Intended level of learning outcomes Lower level learning outcomes: knowledge, recognition, memorization, understanding, without applying what has been learned. Higher level learning outcomes: applying what has been learned (intellectual skills, analysing, synthesising)	1. Lower-level 2. Higher-level 3. Lower + Higher level
Linear	Was the formative CBA linear or adaptive? (If not mentioned in the text, assume it is linear.	1. Linear 2. Adaptive 3. Blocked adaptive 9.Unknown
Methods and procedures		
CgUse	Did the study use a control group? The concept of control group is taken broadly (see next item).	1. Yes 2. No
CgNat	Nature of the control group. Which treatment did the control group receive?	1. Same CBA, no feedback 2. Same CBA, immediate KR only 3. Same CBA, delayed KR only 4. Same CBA, immediate KCR only 5. Same CBA, delayed KCR only 6. Same assessment, but paper-based, no feedback 7. Same assessment, but paper-based, delayed feedback from teacher 8. No assessment 9. Unknown

Assign	Method of assignment to conditions. How were students assigned to the different conditions in the experiment?	1. Random 2. Non-random 3. Random by class 4. By matching, stratification or blocking or by selection on student characteristics (e.g. ability level) 5. Not applicable 6. Other (please specify)
SO2	Please specify other	
Extdes	Experimental design	1. Post-test only 2. Pre-test and post-test (single group) 3. pre-test post-test (multiple groups) 4. non-equivalent groups post-test
EquiG	Was the equivalence of groups tested in a pre-test? Choose not applicable in case there was no pre-test	1. Yes 2. No 3. Not applicable
DifPre	Were there differences in achievement between groups on the pre-test? Choose not applicable in case there was no pre-test	1. No 2. Yes 3. Not applicable
Curri	Were the assessments part of (a course in) the curriculum?	1. Yes, obligatory 2. Yes, not obligatory 3. No 9. Unknown
Summ	Did the post-test have a summative purpose? Did the students receive a grade that counted for their final mark?	1. Yes 0. No 9. Unknown
SupEnv	Did the students complete the post-test assessment in a supervised environment? Did they access the materials in a supervised (class) environment or at home?	1. Yes 0. No 9. Unknown
PoTleng	Length of post-test (number of items)	999 is unknown
PoTrel	Reliability of post-test (α)	999 is unknown

NatPoT	Nature of the post test. How was the learning gain tested? This item is regarding differences with the **pre-test**. Near transfer test means other similar items were used. Far transfer tasks means a totally different type of items were used.	1. Identical test 2. Parallel tests 3. Rewritten items (other form) 4. Different test, near transfer 5. Different test, far transfer 6. Same test items, administered in a different order 999 is unknown
NatPoA	Nature of the post test. How was the learning gain tested? This item is regarding differences with the **assessment for learning** (in which feedback was provided). Near transfer test means other similar items were used. Far transfer tasks means a totally different type of items were used.	1. Identical test 2. Parallel tests 3. Rewritten items (other form) 4. Different test, near transfer 5. Different test, far transfer 6. Same test items, administered in a different order 999 is unknown
PoTatt	Attrition (post-test) % How many participants dropped out during the study?	999 is unknown
Auth	Study authenticity. The study is authentic when the intervention is integrated in the daily practices and curriculum. It is somewhat authentic when it is administered in a classroom situation but stands lose from daily lesson practices. It is not authentic when it is a laboratory experiment	1. Authentic 2. Somewhat authentic 3. Not authentic 4. Unknown
Qual	**Study quality (insufficient – (1) /acceptable +- (2)/ good +(3))** **Selection sample**: 1. Is the sample size per group sufficient to assure the internal validity of the conclusions? <35 is -, > 35 < 90 is +-, > 90 is + 2. Are the characteristics of the sample sufficiently specified ?	1. 2. 3. 4.

	3. Did the authors indicate what the dropout rate was? - is not indicated, +- is indicated but >20%, + is indicated and <20% **Measurement of feedback effects in terms of learning gains:** 4. Is the content of the different feedback conditions clearly defined? 5. Do the authors mention the reliability of the post-test? – is not mentioned, +- is indirectly mentioned, + is mentioned. 6. Is the post-test sufficiently reliable (> .80)? – is < .65, +- is .65 to .79, + is > .80. In case not mentioned (- at 5, choose +-). 7. Do the authors mention the constructs measured by the assessment for learning and the post-test? 8. Do the authors control for an initial student ability or other relevant variables?	5. 6. 7. 8. Judgement: 1. 2. 3. 4. 5. 6. 7. 8. Total:
Effect size information		
Depvar	Dependent variable(s) / construct(s) measured	
TimPoT	Point in time when variable(s) measured; time lag (when was the post-test administered?)	1. Immediately after the formative CBA 2. Shortly (one to three days) after the formative CBA 3. Relatively short (three to seven days) after the formative CBA 4. Between one and three weeks after the formative CBA 5. Other, please specify
S03	Please specify other	
Dattyp	Type of data effect size based on	1. Means and standard deviations 2. T-value 3. F-value 4. Chi-square (df = 1) 5. Other (please specify)
SO4	Please specify other	
SsEg	Sample size experimental group	

SsCg	Sample size control group	
PerFem	Percentage female participants experimental group 999 is unknown, 888 is not applicable	999 888
MageEg	Mean age participants experimental group 999 is unknown, 888 is not applicable	999 888
MageCg	Mean age participants control group 999 is unknown, 888 is not applicable	999 888
MprtCg	Mean pre-test control group 999 is unknown, 888 is not applicable	999 888
MprtEg	Mean pre-test experimental group 999 is unknown, 888 is not applicable	999 888
SdCgPR	Standard deviation pre-test control group 999 is unknown, 888 is not applicable	999 888
ScEgPR	Standard deviation pre-test experimental group, 999 is unknown, 888 is not applicable	999 888
MpotCg	Mean on post-test control group 999 is unknown, 888 is not applicable	999 888
MpotEg	Mean on post-test experimental group 999 is unknown, 888 is not applicable	999 888
SdCgPT	Standard deviation post-test control group	999
ScEgPT	Standard deviation post-test experimental group	999
IVW	Inverse variance weight	
Tval	t-value	
Fval	F-value (df for the numerator must equal 1)	
Chisq	Chi-square (df = 1)	

ProcEs	Procedure used for calculating effect size, including estimation methods	
ES	Effect size statistic	
CrES	Confidence ratings for estimated effect size	
RelES	Reliability of the variables represented in the effect size	
Statt	Statistical technique used in the study (specific for variable)	1. ANOVA or t-test 2. ANCOVA 3. Two-way ANCOVA 4. Multilevel analysis 5. LSD (Least significant differences) 6. General linear model 7. Other
Os5	Please specify other	
Sig	Was the effect for the experimental group positively significant at α = .05?	1. Yes 2. No
PageNr	Page number where the statistics for computing the effect size were found	
SsNote	Study specific notes	

Chapter 5. Interpretation of the Score Reports from the Computer Program LOVS by Teachers, Internal Support Teachers, and Principals[8]

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.

[8] This chapter is a minor revision of Van der Kleij, F. M. & Eggen, T. J. H. M. (2013). Interpretation of the score reports from the Computer Program LOVS by teachers, internal support teachers and principals. *Studies in Educational Evaluation. 39*, 144–152. doi:10.1016/j.stueduc.2013.04.002

5.1 Introduction

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[9] pupil-monitoring system for primary education (LOVS).

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.

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

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 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 programmes 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 amongst 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 of 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 amongst 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 amongst 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 teachers, internal support teachers, and school principals 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).

5.2. Theoretical Framework

5.2.1 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.

5.2.2 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.

5.2.3 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.

5.3. Method

5.3.1 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.

5.3.2 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 experiences 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.

5.3.3 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 analysing 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 ($\alpha = .91$). 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.

5.4. Results

5.4.1 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.

5.4.2 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 5.1, 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 5.1
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

Table 5.1 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 is a significant difference amongst 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 respect 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. Figure 5.1 shows the average proportion correct (P'-value) for each item belonging to a certain report, plotted for each user group.

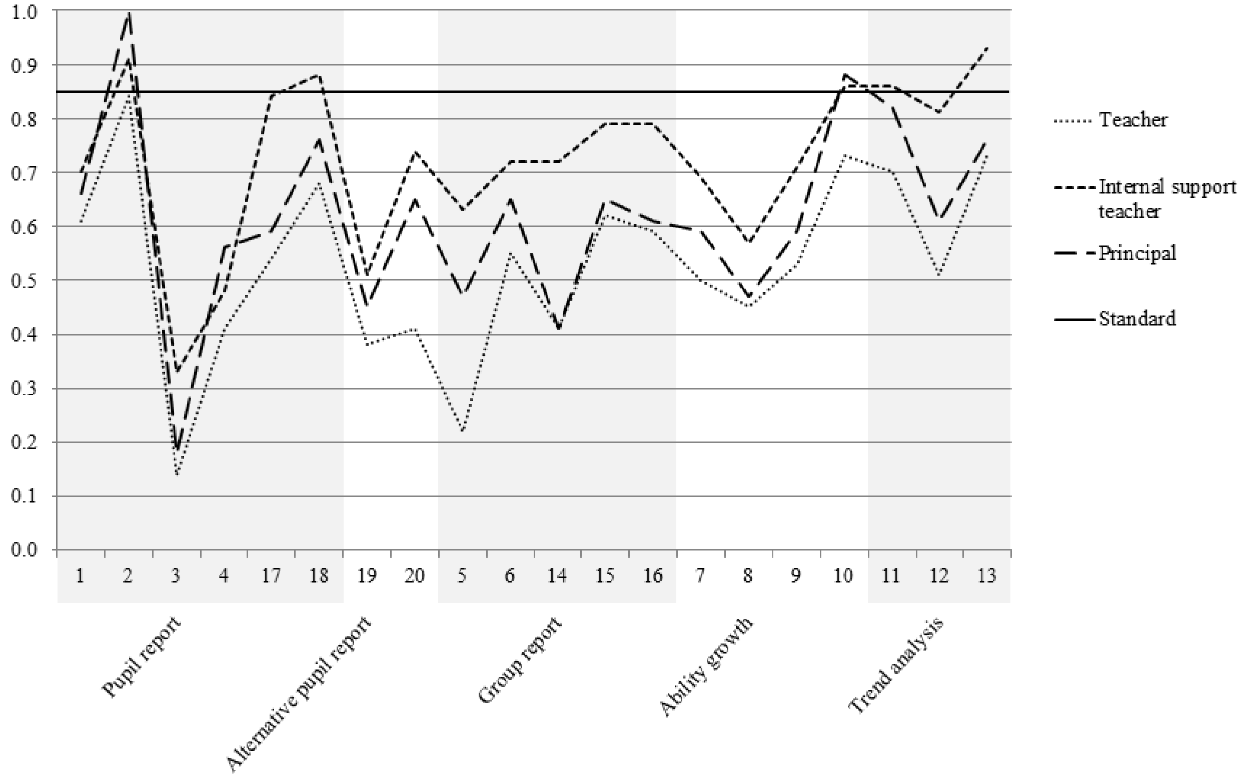


Figure 5.1. Average proportion correct for the three user groups at the item level.

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 report, 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 amongst various user groups at the item level.

From Figure 5.1 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 amongst 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., Item 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 amongst the three groups in terms of the number of respondents who received training. The differences appeared to be large and significant, $F(2, 94) = 19.38$, $p < .001$. 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, $F(1, 95) = 0.71$, $p = .403$.

The number of years' experience using the Computer Program LOVS did not relate to interpretation ability (0–5, 6–10, >10 years) ($F(2, 94) = 1.11$, $p = .331$).

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: $F(2, 93) = 4.82$, $p = .01$.

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, $p = .025$), and between teachers and school principals (average difference = -0.43, $p = .039$). 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: $r(95) = .25$, $p = .013$. 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, $F(2, 94) = 11.64$, $p < .001$. 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, $p < .001$) and school principals (average difference = -0.59, $p = .001$). 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.

5.5. 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 amongst 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 amongst 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 use these 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 Education, Culture, and Science, 2010) and the limited attention that is currently paid to assessment literacy in teacher pre-service programmes. 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, Van Petegem & Vanhoof, 2007; Verhaeghe et al., 2010; Visscher & Coe, 2003; Visscher, & Luyten, 2009; 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, amongst 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).

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. doi: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*, 451–467. doi: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, MA: Harvard Business School.

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

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. doi: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. doi: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. (2009). Visibly learning from reports: The validity of score reports. *Online Educational Research Journal*. Retrieved from http://www.oerj.org/View?action=viewPDF&paper=6

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. doi:10.2190/ET.36.2.g

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112. doi:10.3102/003465430298487

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. Retrieved from http://ld6ela.edublogs.org/files/2008/07/data-article-Kerr-et-al.pdf

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, the Netherlands: SCO-Kohnstamm Instituut.

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, UK: Longman.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice*, *48*, 4–11. doi: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. doi: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. doi: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. doi: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. doi: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]. Arnhem, the Netherlands: Cito.

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*(5)*,* 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. doi:10.1080/13803610701702522

Vanhoof, J., Van Petegem, P., & De Maeyer, S. (2009). Attitudes towards school self-evaluation. *Studies in Educational Evaluation, 35*, 21–28. doi:10.1016/j.stueduc.2009.01.004

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. doi:10.1080/03055698.2010.482771

Verhaeghe, G. (2011). *School performance feedback systems: Design and implementation issues.* (Doctoral dissertation, University of Gent, Belgium). Retrieved from http://users.ugent.be/~mvalcke/CV/PhD%20Goedele%20Verhaeghe.pdf

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. doi: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. doi:10.1016/S0191-491X(01)00026-8

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, the Netherlands: Swets & Zeitlinger.

Visscher, A. J., & Coe, R. (Eds.) (2002). *School improvement through performance feedback*. Lisse, the Netherlands: 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. doi: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. doi: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. doi: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. doi:10.1080/09243450802696695

Chapter 6. Towards Valid Score Reports in the Computer Program LOVS: A Redesign Study[10]

Abstract

Recent research in the field of education (Van der Kleij & Eggen, 2013) has suggested that users do not make an entirely correct interpretation of the score reports from the pupil-monitoring Computer Program LOVS. This study investigated how these reports can be redesigned in a way that supports users in interpreting pupils' test results. In several rounds of consultations and designs with the users and experts, alternative designs for the reports were created and field tested. No clear differences were found in terms of users' interpretation accuracy between the original and the redesigned versions of the reports. However, users' perceptions of the redesigned reports were predominantly positive. Eventually, a final set of design solutions was generated. The authors emphasise the need for clear score reports and involvement of experts and current/future users in the design process to ensure the validity of the reports.

[10] This chapter has been submitted as Van der Kleij, F. M. Eggen, T. J. H. M., & Engelen, R. J. H. (submitted). *Towards valid score reports in the Computer Program LOVS: A redesign study.* Manuscript submitted for publication.

6.1. Introduction and Study Context

Recently, Van der Kleij and Eggen (2013) investigated the interpretation of the score reports from the Computer Program LOVS, Cito's (the Institute for Educational Measurement in the Netherlands) pupil-monitoring system for primary education. The results suggested that users, particularly teachers, do not make an entirely correct interpretation of these reports, and they encounter many stumbling blocks. This situation is problematic, given the demands placed on all actors within Dutch schools with respect to the implementation of data-driven decision making (DDDM) (Ministry of Education, Culture, & Science, 2010). Sometimes called data-based decision making (e.g., Schildkamp, Lai, & Earl, 2013), DDDM has been increasingly popular, as it is considered a promising means of improving pupils' learning outcomes. However, a correct interpretation of data regarding student learning is a necessary precondition for DDDM to fulfil its potential.

Pupil-monitoring systems provide users with feedback about pupil performance in the form of a score report at the level of the individual pupil, class, or school; this is called data feedback. The data feedback is intended to inform learning in order to close the gap between a pupil's current performance and the intended learning outcomes (Hattie & Timperley, 2007; Sadler, 1989). Thus, score reports from pupil-monitoring systems aim to steer future educational decisions and activities at various levels within the school, which is at the heart of DDDM.

6.1.1 The Pupil-Monitoring Computer Program LOVS

The pupil-monitoring system LOVS consists of a coherent set of nationally standardised tests for longitudinal assessment of primary school children. Covering skills such as reading comprehension, spelling, and math, the LOVS tests are usually administered twice a year.

Various computer programs are available for registering and processing the test results, but this study solely focuses on Cito's Computer Program LOVS. This computer program allows the automatic generation of score reports at the level of the pupil, group, or school. The reports provide both graphical representations and numerical information, either in a table or a graph, or a combination of both. In all LOVS tests, the items are calibrated onto an Item Response Theory (IRT) scale. Every subject area has its own unique scale, which implies that the pupils' abilities cannot be directly interpreted. In order to give meaning to the pupils' abilities, level indicators that represent certain percentile scores in the population are used. At the levels of the individual pupil and the class, the Computer Program LOVS aims to feed back information about student learning for instructional improvement, which is considered a formative purpose (Stobart, 2008). At the levels of the class and the school, the reports are intended for the school's own evaluation purposes, both formatively and summatively.

6.1.2 DDDM Using LOVS Tests

The DDDM approach encompasses a systematic and cyclic process. In this study, we used the cycle of the pupil-monitoring system LOVS, which consists of the signaling, analysis, and acting phases (Figure 6.1). In the first phase, the test is administered, and the results are checked, whereupon the test results are registered and interpreted by users. Subsequently, when needed, users seek additional information in the analysis phase to find possible explanations for the test scores. This phase results in a plan for future action. Next, the plan is executed; eventually, the effects of the intervention are evaluated. When the three phases have been completed, the cyclic process starts all over again. In a subsequent cycle, users can evaluate whether or not the intervention was effective. This study is aimed at interpreting the test results, which is part of the signaling phase.

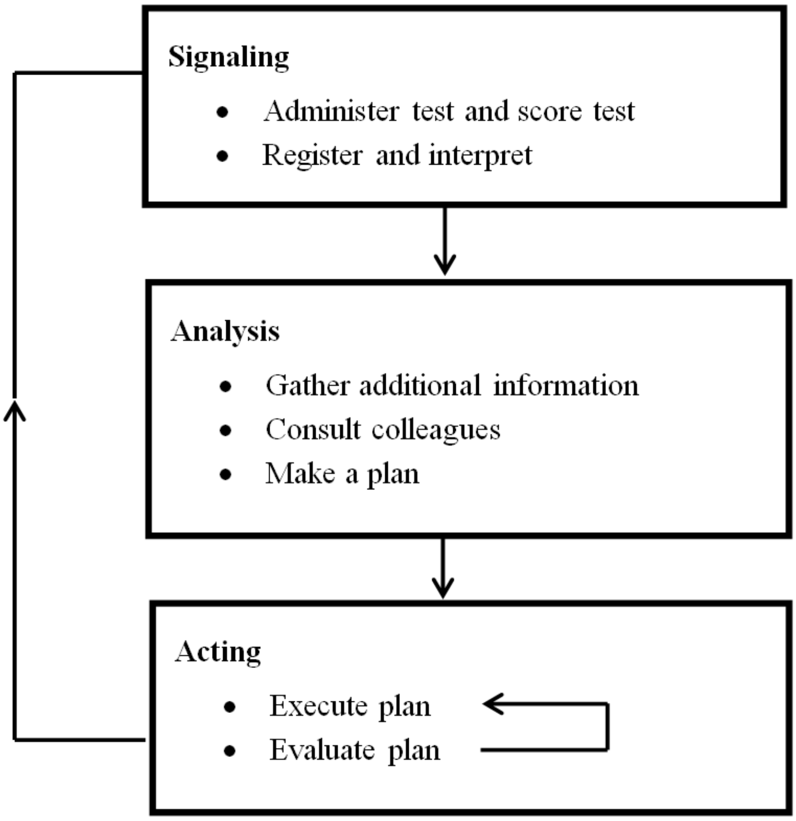


Figure 6.1. The evaluative cycle of the pupil-monitoring system LOVS.

6.1.3 Interpretability as Aspect of Validity

Supporting users in interpreting pupils' assessment score reports has recently been addressed as an important aspect of validity (Hattie, 2009; Ryan, 2006; Zenisky & Hambleton, 2012). The validity of the test scores is negatively affected whenever test results are not interpreted as the test developer intended (Hattie, 2009; Ryan, 2006). A correct interpretation of reports is especially relevant when the test results are meant to inform important or irreversible decisions. Although this is usually not the case for formative uses of assessment results (Bennett, 2011), a correct interpretation is a precondition for tailoring instruction or adapting the learning environment in the broader sense to the needs of all the pupils in order to realise their learning potential.

6.1.4 The Need for Professional Development in DDDM

In the last few years, it has become increasingly clear that implementing DDDM is a challenging undertaking. Moreover, various professional development initiatives have been established that aim to make school staff proficient in using data from pupil-monitoring systems (e.g., Staman, Visscher, & Luyten 2013). The idea behind such initiatives is that a certain degree of "assessment literacy" is needed for interpreting test results (Earl & Fullan, 2003; Vanhoof, Verhaeghe, Verhaeghe, Valcke, & Van Petegem, 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). From this definition, it can be reasoned that the lack of an accurate understanding of data about student learning directly affects a person's subsequent actions. Thus, a correct interpretation of data about student learning is a necessary precondition for successfully implementing DDDM.

Professional development on the interpretation and use of pupil-monitoring system data seems necessary in the Dutch context, especially given the lack of formal requirements in past and current pre-service teacher education programmes (Van der Kleij & Eggen, 2013). Results from recent research (Staman et al., 2013), for example, show that school staff can benefit from a long-term, intensive, school-wide training programme. Nevertheless, it appears that even after such a rigorous programme, users' interpretations are not free of errors. Conversely, clear score reports can contribute to the accuracy of users' interpretations. Moreover, test developers are to a certain extent responsible for providing score reports that support users in making correct interpretations (Hattie, 2009; Ryan, 2006; Zenisky & Hambleton, 2012).

6.1.5 Aims of the Present Study

This study aims to set an example for design research (McKenney & Reeves, 2012) in the area of score reporting. Research has been conducted in the context of the reports generated by the Computer Program LOVS. This study focused on five reports, two at the pupil level, and three at the group level. The reports at the pupil level are to be used to monitor individual progress and to signal abnormalities. However, some abnormalities for example, stagnation in the learning curve, are not explained by the reports, but will have to be examined in the analysis phase. For this purpose, additional reports that allow for specific error analyses are available. The three reports at the group level are intended to be used for internal evaluation purposes at the levels of the class and/or the school.

This study investigated how the reports from the Computer Program LOVS can be redesigned to support users in interpreting pupils' test results. The aims of this study were twofold, as is typical for design research (McKenney & Reeves, 2012). First, solve a problem in practice, i.e., users, particularly teachers, seem to experience difficulties in interpreting the reports generated by the Computer Program LOVS. Second, contribute to the theoretical understanding regarding score reporting.

6.2. Theoretical Framework

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 how to report test scores. These standards are not directly usable for the (re)design process of score reports. In recent years, several studies have attempted to provide guidelines for improving score reporting practices (e.g., Goodman & Hambleton, 2004; Hattie, 2009; Ryan, 2006; Zenisky & Hambleton, 2012), which can be very useful for direct application.

Zenisky and Hambleton (2012) have summarised the literature on score reporting in terms of best practices. They have outlined the recommendations based on research in five areas: 1) report development processes, 2) report design and format, 3) contents, 4) ancillary materials, and 5) dissemination efforts. This study focused on the first three areas, although the authors acknowledge the importance of the fourth and fifth areas as well.

From the literature about score reports in educational and multimedia research, various design principles can be extracted regarding the content, design, and format (many of which have been summarised by Zenisky & Hambleton, 2012). Furthermore, Kosslyn (2006) has distinguished eight principles for the design of visual displays, inspired by various research disciplines, such as psychology. For this study, the principles from the educational literature were integrated into Kosslyn's (2006) eight principles:

1. *Principle of Relevance.* The graph designer should bear in mind the particular message the graph needs to communicate. Moreover, score reports should be designed for a specific test purpose (Zenisky & Hambleton, 2012). Ideally, there is a good balance between presenting essential information and details. According to Mayer (2001), irrelevant information should be avoided whenever possible, since it causes unnecessarily high cognitive load and distraction. A score report should contain all information necessary for interpretation, as intended by the designer. This includes a description of the test purpose, intended uses for the test and test scores, and confidence intervals with accompanying explanations (Goodman & Hambleton, 2004; Ryan, 2003). Furthermore, various strategies can be used to support interpretation, such as anchor points, market-basket displays, and benchmarking (Jaeger, 2003; Mislevy, 1998). These methods will help the user understand the test results in terms of both content-referenced and criterion-referenced interpretations (Zenisky & Hambleton, 2012). Reporting subscores can help indicate strengths and weaknesses, and improve the formative potential of the score report. However, from a psychometric viewpoint, reporting at a fine-grained level (such as attribute subscores) can raise reliability issues, depending on the properties of the test items (Monaghan, 2006). Ryan (2003) has suggested reporting results at the finest level possible, at an acceptable level of accuracy. What is acceptable in terms of reliability depends largely on the degree of reversibility of the intended decisions based on the test results (Nitko & Brookhart, 2007). Furthermore, Goodman and Hambleton (2004) have suggested that a certain degree of personalisation (e.g., inserting the pupil's name) can help the user connect with the score report.

2. *Principle of Appropriate Knowledge.* It is advised that the target user(s)'s characteristics be taken into account. This approach counts for both the selected type of graph and the data presented in it. Research suggests that users prefer types of reports similar to what they are accustomed (Wainer, Hambleton, & Meara, 1999). Moreover, Kosslyn (2006) has emphasised that a display should build on knowledge that the reader already possesses. Furthermore, a list explaining the key terms used in the report can be helpful (Tufte, 1983, 1990; Wainer, 1997). Various researchers have discouraged the use of statistical jargon (Goodman & Hambleton, 2004; Hambleton & Meara, 2000; Ryan, 2003; Tufte, 1983, 1990; Wainer, 1997). The use of decimals has also been dissuaded (Tufte, 1983, 1990; Wainer, 1997).
3. *Principle of Salience.* The most important information should be highlighted (Goodman & Hambleton, 2004; Kosslyn, 2006; Ryan, 2003, Tufte, 1983, 1990; Wainer, 1997) and thus presented to attract attention, for example, by using boldface text (Kosslyn, 2006), frames, or visual displays (Tufte, 1983, 1990; Wainer, 1997). However, Kosslyn has emphasised that what is "visually striking" depends on all properties of the display. In other words, how much attention a certain aspect will draw is always relative. Making the score reports as clear as possible by avoiding clutter (Goodman & Hambleton, 2004; Tufte, 1983, 1990; Wainer, 1997), and using sufficient white space (Leeson, 2006) have also been proposed. Moreover, a score report should be "actionable" (Zenisky & Hambleton, 2012), i.e., suggest a future course of action in the learning process (Goodman & Hambleton, 2004; Hattie, 2009).
4. *Principle of Discriminability.* The graph properties should be sufficiently different from one another to be distinguishable by the user (Kosslyn, 2006). For example, two lines representing two groups should be separately distinctive.
5. *Principle of Perceptual Organisation.* People will automatically group elements into patterns, which is influenced by their colour and ordering. For example, xxxx is seen as a cluster of elements, but xx xx is viewed as two groups. Furthermore, it is advised that the data be grouped in a meaningful way (Goodman & Hambleton, 2004). Using bar graphs is recommended for comparison purposes, where individual bars can best be arranged by height (Tufte, 1983, 1990; Wainer, 1997). Colours can also be used in meaningful ways (Tufte, 1983, 1990; Wainer, 1997), e.g., cool colours at the background and warm colours at the foreground, because warm tones appear closer to the human eye than cool hues (Kosslyn, 2006). According to Leeson (2006), people are likely to read a paper report, but are inclined to "scan" reports displayed onscreen. Wainer et al. (1999) have also claimed that reports have to be seen, not merely read. Thus, this is possibly an important aspect to bear in mind when (re)designing digital score reports.
6. *Principle of Compatibility.* The shape of the message should be compatible with the structure of the graph. This implies that something displayed as 'more' in the graph must visually match 'more' of something. For example, a higher bar in a bar chart would mean a higher test score. Also, certain patterns and conventions should be considered. In Western cultures, for example, the colour red means 'stop' and green signifies 'go'. Ignoring the meanings of these colours in certain cultures violates this

principle (Kosslyn, 2006). Moreover, as previously mentioned, colours should be used in meaningful ways (Tufte, 1983, 1990; Wainer, 1997).

7. *Principle of Informative Changes.* People expect certain changes to contain information. For example, a rising line in a graph means an increase of something. Furthermore, it is important to label all relevant aspects of the graph (Kosslyn, 2006).
8. *Principle of Capacity Limitations.* Human beings are capable of processing a limited amount of information at one instance. Therefore, the quantity of data presented should be restricted. It has been advised that a combination of displays or graphs and supporting text be employed whenever possible (Goodman & Hambleton, 2004; Hattie, 2009; Leeson, 2006; Tufte, 1983, 1990; Wainer, 1997). By using both visual and textual forms of representation, information can be processed more easily. This way, the reader gets the chance to create both verbal and graphical mental models, and to link these models to one another. Visual and textual representations should be placed at a close proximity to one another on the screen or page. This increases the likelihood for both representation forms to remain in working memory, which facilitates effective processing (Mayer, 2001).

The foregoing design principles are of a general nature and may have different consequences, depending on the level of reporting of the test results and the proposed uses of the test. According to Ryan (2006), the reporting unit is one of the key characteristics to consider in designing a score report. In many tests, the results are fed back to the test taker and other stakeholders (e.g., the teacher and parents). These score reports are aimed at the individual level. In various testing programmes, the results are also aggregated to a higher level (e.g., the class or the school) and reported accordingly (Zenisky & Hambleton, 2012). This has implications for reports; multiple reports are often needed to communicate the results of one test at various levels.

The importance of the degree in which the report is actionable depends on the proposed use of the test. Some reports are mainly aimed at communicating a pass/fail decision corresponding to a summative test purpose, while score reports of tests with a formative purpose should especially provide suggestions for future learning and/or modifications in instruction (Bennett, 2011).

6.3. Design Research Approach

This study followed an educational design research approach (McKenney & Reeves, 2012). McKenney and Reeves have distinguished various phases within the design research process: 1) analysis and exploration, 2) design and construction, and 3) evaluation and reflection. This process eventually results in two outcomes: An intervention and theoretical understanding. The process is iterative and flexible, and takes place within the intended context of the intervention. In this study, several cycles of design and construction were followed up by (micro-)evaluations and a subsequent redesign.

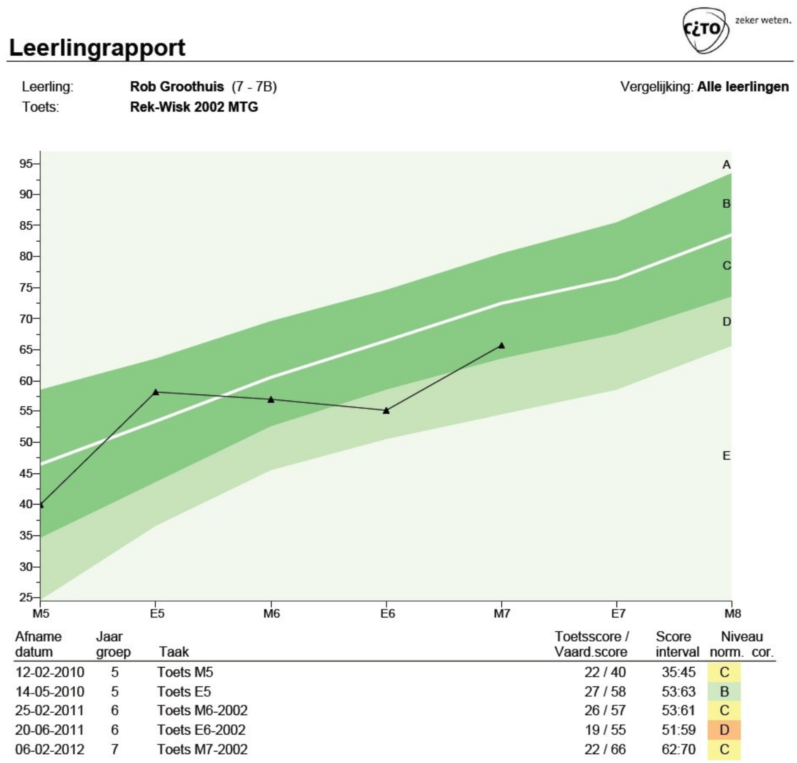
This study used the approach characterised by McKenney and Reeves (2012) as research *on interventions*. The intervention that is the subject of this investigation comprises the reports generated by the Computer Program LOVS. Thus, the solution to the problem at hand would be sought in the product form of redesigned score reports. Furthermore, an existing theory in the form of design principles for score reports was used as the basis for the design. The results of this investigation aimed to contribute to the theoretical understanding about score report interpretation.

Various researchers have stressed the importance of field testing and conducting focus group meetings when (re)designing the contents of score reports (Allalouf, 2007; Hambleton & Slater, 1997; Hattie, 2009; Trout & Hyde, 2006; Wainer et al., 1999). Hattie proposed using proficiency tests, focus group meetings, and insights from previous research as sources of information regarding the adequacy of users' interpretation of score reports. He emphasised the scarcity of empirical studies that have accounted for actual user interpretation. Zenisky and Hambleton (2012) also highlighted the need for studies that investigate the actual understanding of score reports. A design research approach is suitable to the problem under review, because it allows for detailed investigations in the intended context of the intervention in collaboration with users.

The first author was present in each of the sessions with the focus groups and key informants. In the focus group meetings, the researcher was accompanied by an educational adviser. Each session was supported by a PowerPoint presentation (structured based on the five reports), and was digitally audiotaped, transcribed, and summarised. For the focus group meetings, the contents of the reports were verified by the contact persons in the schools, i.e., member checking (Creswell & Plano Clark, 2007).

6.4. Analysis and Exploration

The analysis and exploration phase usually consists of an initial orientation, a literature study, and an investigation within the field. Van der Kleij and Eggen (2013) have already explored the problem under investigation. Experts were consulted, focus group meetings at two schools were held, and a group of users of the Computer Program LOVS filled out a questionnaire that measured their ability to interpret the reports. The present study focused on the same five reports generated by the Computer Program LOVS: The pupil report (Figure 6.2), the alternative pupil report (Figure 6.3), the group report (one test-taking moment) (Figure 6.4), ability growth (Figure 6.5), and trend analysis (Figure 6.6). These reports have been most frequently used in schools or have often been interpreted incorrectly. The questionnaire data gathered by Van der Kleij and Eggen were used as a pre-test measure for this redesign study. The same focus groups agreed to participate and collaborate for the purpose of redesigning the reports. Moreover, the sample of 56 schools drawn by Van der Kleij and Eggen (2013) was used for administering a questionnaire to evaluate the redesigned reports. An additional literature study was conducted to identify design principles for score reports that are reported in the conceptual framework.



Leerlingrapport

zeker weten.

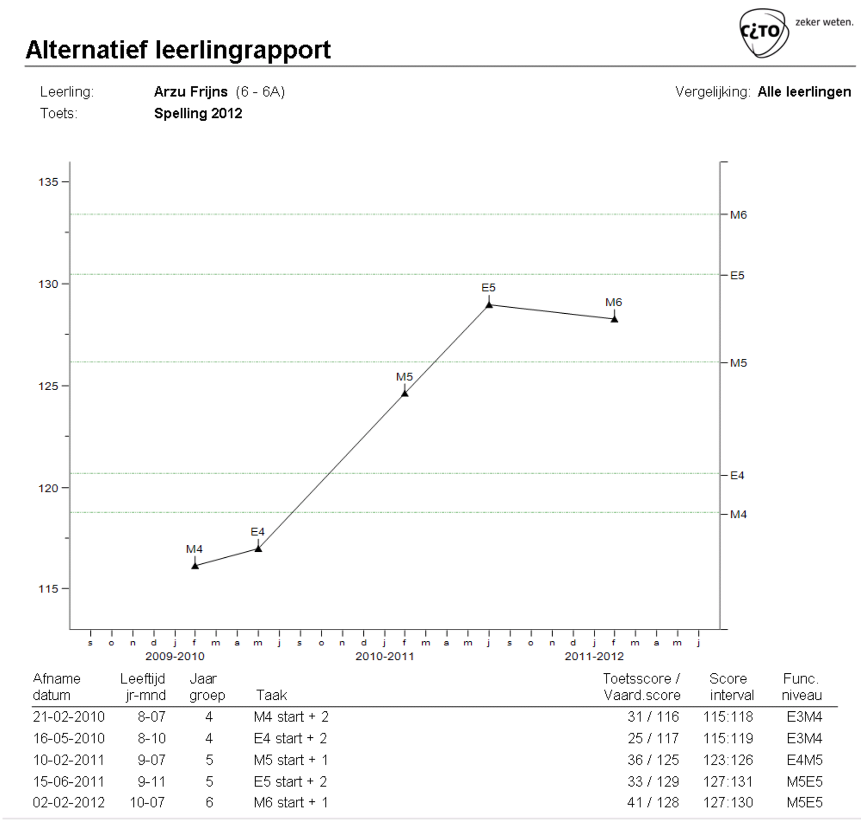
Leerling: **Rob Groothuis** (7 - 7B)

Toets: **Rek-Wisk 2002 MTG**

Vergelijking: **Alle leerlingen**

Afname datum	Jaar groep	Taak	Toetsscore / Vaard.score	Score interval	Niveau norm.	cor.
12-02-2010	5	Toets M5	22 / 40	35:45	C	
14-05-2010	5	Toets E5	27 / 58	53:63	B	
25-02-2011	6	Toets M6-2002	26 / 57	53:61	C	
20-06-2011	6	Toets E6-2002	19 / 55	51:59	D	
06-02-2012	7	Toets M7-2002	22 / 66	62:70	C	

Figure 6.2. Pupil report.



Alternatief leerlingrapport

zeker weten.

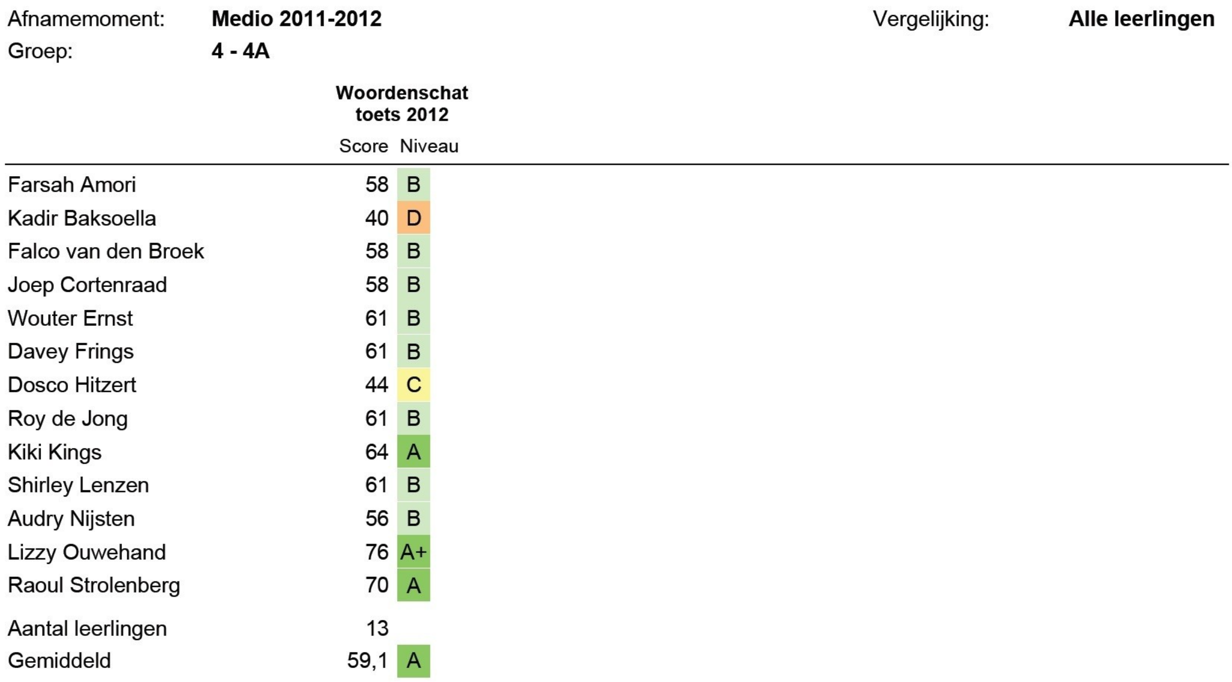
Leerling: **Arzu Frijns** (6 - 6A)

Toets: **Spelling 2012**

Vergelijking: **Alle leerlingen**

Afname datum	Leeftijd jr-mnd	Jaar groep	Taak	Toetsscore / Vaard.score	Score interval	Func. niveau
21-02-2010	8-07	4	M4 start + 2	31 / 116	115:118	E3M4
16-05-2010	8-10	4	E4 start + 2	25 / 117	115:119	E3M4
10-02-2011	9-07	5	M5 start + 1	36 / 125	123:126	E4M5
15-06-2011	9-11	5	E5 start + 2	33 / 129	127:131	M5E5
02-02-2012	10-07	6	M6 start + 1	41 / 128	127:130	M5E5

Figure 6.3. Alternative pupil report.



Afnamemoment:	Medio 2011-2012	Vergelijking:	Alle leerlingen
Groep:	4 - 4A		

	Woordenschat toets 2012	
	Score	Niveau
Farsah Amori	58	B
Kadir Baksoella	40	D
Falco van den Broek	58	B
Joep Cortenraad	58	B
Wouter Ernst	61	B
Davey Frings	61	B
Dosco Hitzert	44	C
Roy de Jong	61	B
Kiki Kings	64	A
Shirley Lenzen	61	B
Audry Nijsten	56	B
Lizzy Ouwehand	76	A+
Raoul Strolenberg	70	A
Aantal leerlingen	13	
Gemiddeld	59,1	A

Figure 6.4. Group report.

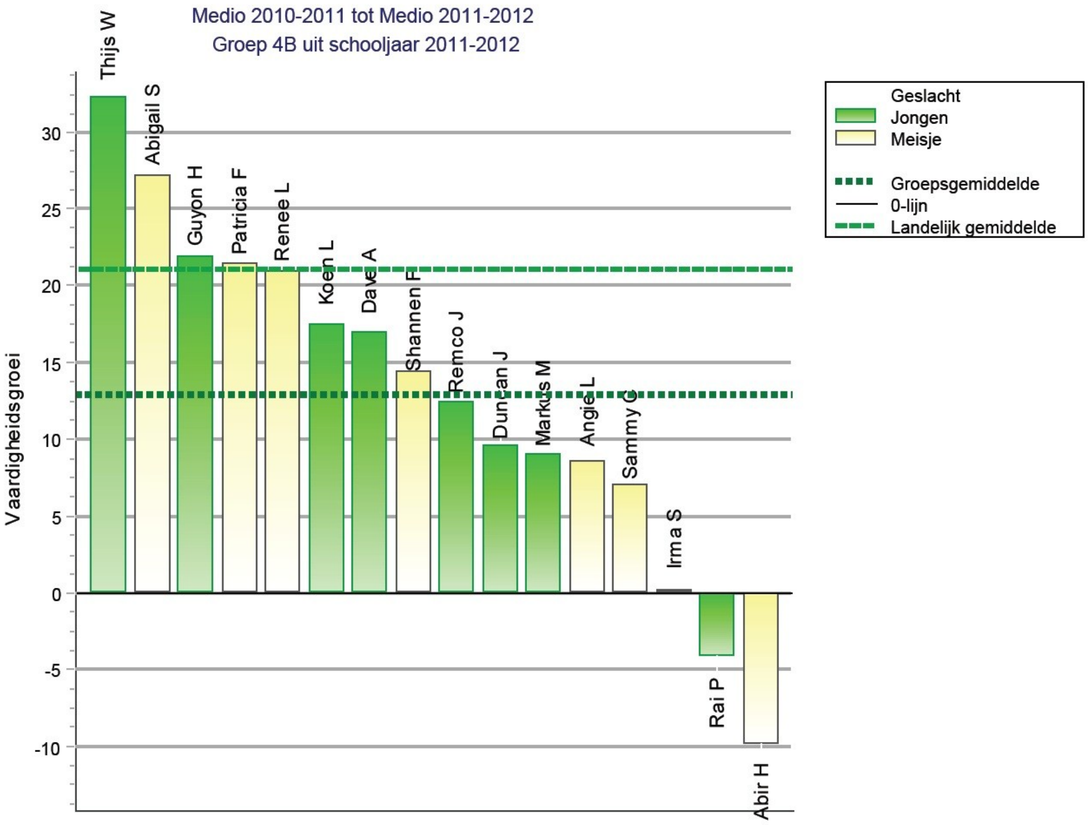


Figure 6.5. Ability growth.

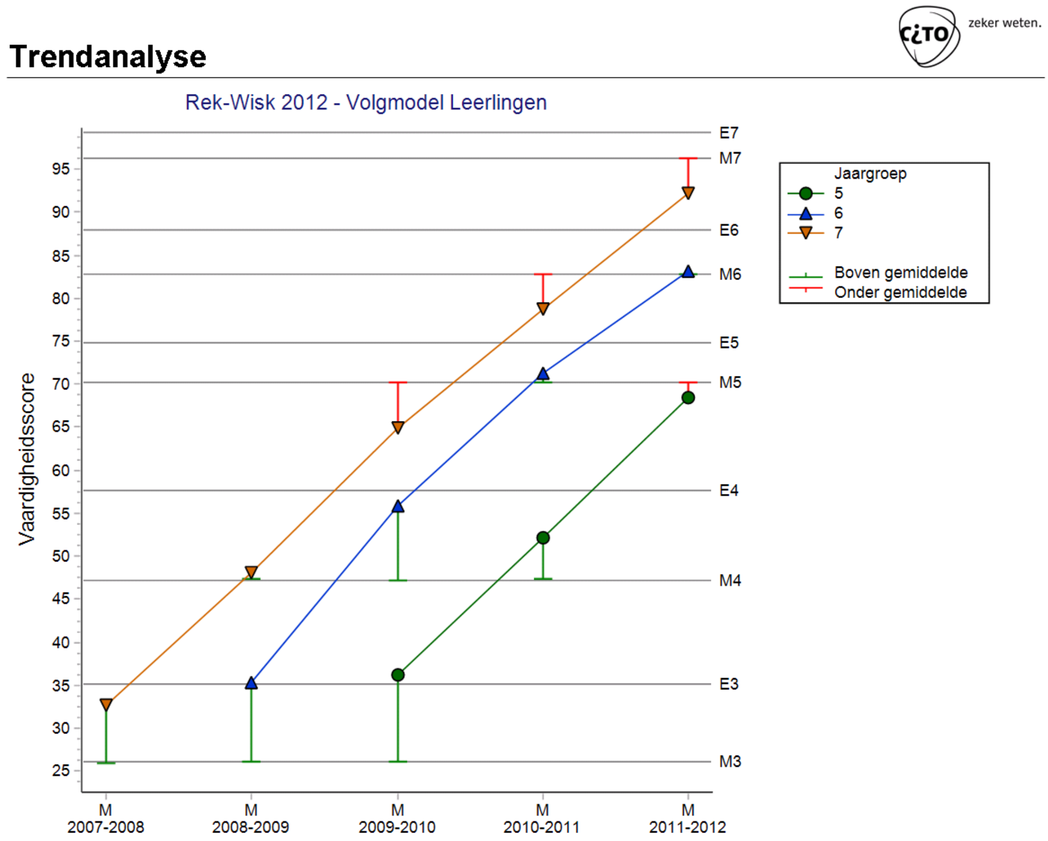


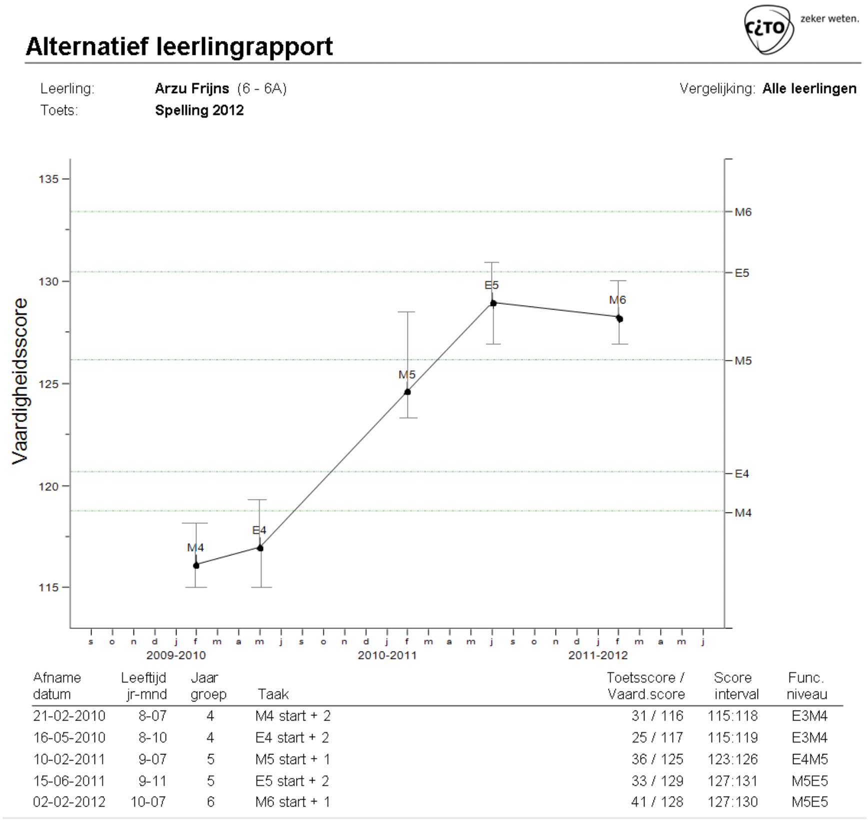
Figure 6.6. Trend analysis.

6.5. Design and Construction

6.5.1 Initial Design

Van der Kleij and Eggen (2013) distinguished several knowledge aspects and interpretation skills required to correctly interpret the reports from the Computer Program LOVS. For example, users should know the meanings of the level indicators (A–E and I–V) and the score intervals, and interpret ability growth as opposed to ability. Based on the outcomes of the focus group meetings and questionnaire by Van der Kleij and Eggen, prototypes for alternative score reports were created using the design principles.

The original reports were used as a starting point for the prototypes, based on the suggestion that users prefer types of reports similar to those with which they are familiar (Wainer et al., 1999). Furthermore, according to the experts, many users were satisfied about the formats of the original reports. For each report, multiple prototypes were generated by the first author, which served as design drafts. The prototypes were formatively evaluated in consultation with the experts and the two focus groups. For each report, the original report was graphically displayed, followed by a summary of the results from the questionnaire and focus group meetings (Van der Kleij & Eggen, 2013), highlighting the aspects in need of redesign. Subsequently, multiple prototypes resulting from several possible design solutions were presented for each report. Figure 6.7 shows a sample prototype for the alternative pupil report.



Alternatief leerlingrapport

Leerling: **Arzu Frijns** (6 - 6A) Vergelijking: **Alle leerlingen**

Toets: **Spelling 2012**

Afname datum	Leeftijd jr-mnd	Jaar groep	Taak	Toetsscore / Vaard.score	Score interval	Func. niveau
21-02-2010	8-07	4	M4 start + 2	31 / 116	115:118	E3M4
16-05-2010	8-10	4	E4 start + 2	25 / 117	115:119	E3M4
10-02-2011	9-07	5	M5 start + 1	36 / 125	123:126	E4M5
15-06-2011	9-11	5	E5 start + 2	33 / 129	127:131	M5E5
02-02-2012	10-07	6	M6 start + 1	41 / 128	127:130	M5E5

Figure 6.7. Prototype for alternative pupil report.

6.5.2 Expert Evaluation

Three experts were asked to comment on the prototype aspects relevant to an easier and more accurate interpretation. The experts listed their comments in a document, which were discussed and where necessary, clarified in a telephone conversation or by e-mail.

6.5.3 Focus Groups

The researchers visited the two schools and discussed the prototypes in focus group meetings. The focus groups were set up in a group discussion format (Newby, 2010). In School 1, four teachers (one of whom was doing an internship), one internal support teacher, and two principals participated. In School 2, two teachers, one ICT teacher/coordinator, and one administrative member participated. Qualitative data were gathered about the participants' perceptions of the prototypes with respect to ease of interpretation.

6.5.4 Proposed Design

Based on the expert evaluations and focus group meetings, revisions were made to the design solutions, and a preliminary design was decided upon. A graphic artist created the preliminary designs of the reports. Table 6.1 shows the aspects in need of redesign (see Van der Kleij & Eggen, 2013), the relevant design principles from the literature, the proposed design solution, and the report to which the change was applied. The principle of capacity limitations was not used for the purpose of redesign, because there were no indications of users being overwhelmed by the contents of the reports.

Table 6.1

Aspects in Need of Redesign, Design Principle, and Design Solution(s) for Each Report

Aspects in need of redesign	Design principle	Design solution	Report
Knowledge skills			
Know the meaning of the level indicators (A–E and I–V) and know the position of the national average within the different levels.	• Relevance	• Add a legend showing the meanings of the level indicators, including a dotted line for the national average.	• Pupil • Group
Know the meaning of the score intervals around the ability.	• Salience	• Draw the score intervals in the graph.	• Pupil • Alternative pupil • Group
Know that the norms for groups differ from those for individual pupils with regard to the level indicators.	• Relevance • Salience • Discriminability	• Remove irrelevant information: “comparison all pupils”. • Frame the level of the group. • Add “comparison all schools” and frame the level of the group.	• Group
Interpretation skills			
Understand to which level a trend is referring.	• Salience	• Boldface horizontal lines in case of a yearly comparison.	• Trend analysis
Interpret ability growth as opposed to ability scores.	• Salience	• Display ability level indicators in bars that display growth in ability.	• Pupil • Alternative pupil • Ability growth
Understand whether the growth of the group is under or above the national average.	• Discriminability • Perceptual organisation	• Change style and colour of the lines referring to the group average and national average.	• Ability growth
Understand whether the level of the group is under or above the national average.	• Appropriate knowledge	• Add a legend showing the meanings of the level indicators and a dotted line for the national average.	• Group
Other (design)			
Confusing use of symbols.	• Compatibility	• Change triangles into other symbols.	• Pupil • Alternative pupil • Trend analysis
Confusing use of colour.	• Compatibility	• Avoid using the colours green, red, and dark orange to indicate groups. • Use new colour coding for level indicators (I-V as default). • Adapt colours in graphs to the colours of the particular levels (as in table).	• Pupil • Group • Trend analysis
Notation of score intervals in table.	• Compatibility	• Substitute “:” with “-” to indicate a range as opposed to a ratio.	• Pupil • Alternative pupil
Label of y-axis is lacking.	• Relevance • Informative changes	• Add appropriate labels to the axes.	• Pupil • Alternative pupil
Difference between school years and display of test-taking moments is unclear.	• Salience	• Place vertical lines between school years and boldface test-taking moments.	• Alternative pupil

6.6. Evaluation and Reflection

This section describes the evaluation and reflection phase. The researchers want to emphasise, however, that the design and evaluation phases occurred in iterative cycles and was not a linear process.

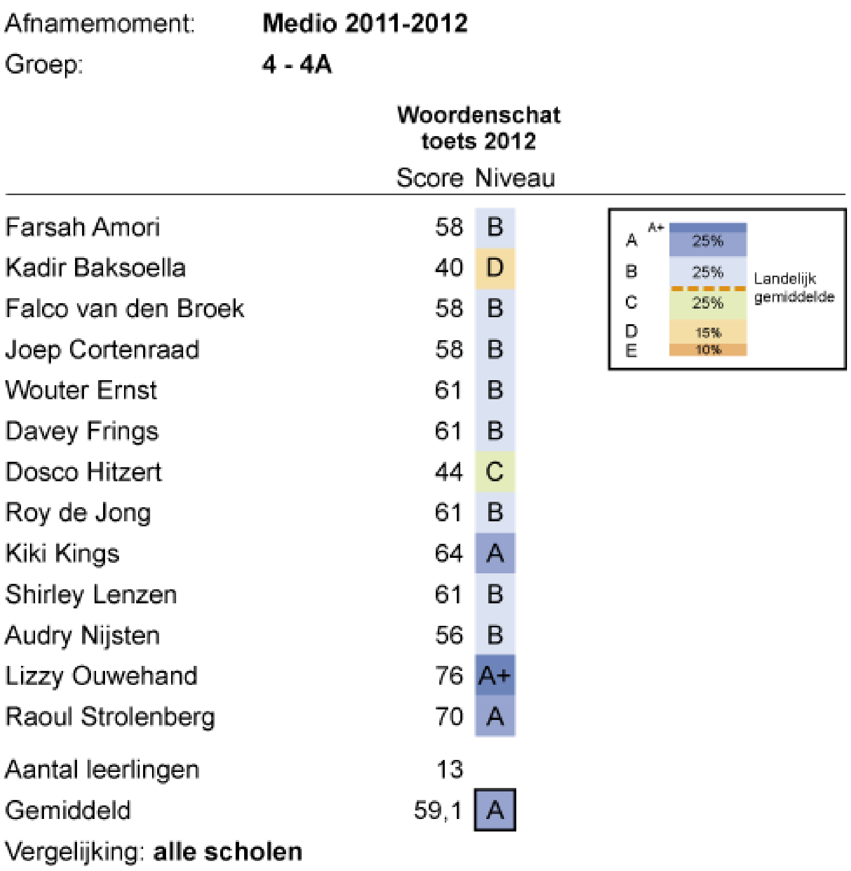
The preliminary designs of the reports were evaluated in a questionnaire, which served as a post-test. It was expected that the redesigned reports would be easier to interpret than the original ones; therefore, the respondents were asked to indicate their opinion on this issue. Furthermore, the redesigned reports were evaluated in consultation with the two focus groups. These evaluations not only served to identify how the designs could be further improved, but also had to indicate how effective the intervention was in terms of interpretation accuracy. Subsequently, the preliminary designs were adapted as needed, whereupon some of the easily adaptable revisions were implemented. Finally, key informants were consulted to gather detailed feedback on the proposed and implemented design solutions. Eventually, a final design solution was proposed.

6.6.1 Questionnaire

Instruments and procedure. The online questionnaire used by Van der Kleij and Eggen (2013) was adapted for the evaluation of users' interpretation of the redesigned reports. The items for the current questionnaire were chosen based on both content and psychometric grounds. For example, items that had very high P'-values (average proportion correct) were excluded because they were not informative. Furthermore, images of the reports were replaced by images of their redesigned versions (see Figure 6.8 for an example). The contents of the items remained unchanged. However, it was necessary to change the scoring for one item, given the adapted image of the report.

Two versions of a 30-item questionnaire were produced, each containing different anchor items that belonged to an image of the original report. This was needed to compare the results between the original and redesigned reports. For measuring user ability, 13 items were used, which had either a multiple-choice or a multiple-response format. These items were scored as 34 separate dichotomous items. Also, 10 items concerned the respondents' background characteristics, such as their function within the school, their years of experience with the Computer Program LOVS, and whether or not they had received training in the use of the Computer Program. Furthermore, the questionnaires contained six items asking the respondents to indicate whether they thought that the redesigned reports were easier to interpret (with the original and redesigned versions displayed next to each other). For these items, a five-point Likert scale was used, ranging from totally disagree to totally agree. In addition, one open item allowed the respondents to leave comments.

The questionnaires could be filled out during a two-week period. Respondents could leave their e-mail addresses if they wanted to receive elaborated feedback on how to correctly interpret the reports after the closing date of the questionnaire.



Afnamemoment: **Medio 2011-2012**

Groep: **4 - 4A**

	Woordenschat toets 2012	
	Score	Niveau
Farsah Amori	58	B
Kadir Baksoella	40	D
Falco van den Broek	58	B
Joep Cortenraad	58	B
Wouter Ernst	61	B
Davey Frings	61	B
Dosco Hitzert	44	C
Roy de Jong	61	B
Kiki Kings	64	A
Shirley Lenzen	61	B
Audry Nijsten	56	B
Lizzy Ouwehand	76	A+
Raoul Strolenberg	70	A
Aantal leerlingen	13	
Gemiddeld	59,1	A

Vergelijking: **alle scholen**

Figure 6.8. Redesigned draft group report.

Respondents. The contact persons in the 56 schools that participated in the Van der Kleij and Eggen (2013) study were reached by e-mail and asked to disseminate the questionnaire about the redesigned reports in their schools. A call for participants was also posted on the Cito website, asking users to fill out the online questionnaire. In total, 74 unique and sufficiently complete responses were retrieved, of which 36 had filled out questionnaire 2, version 1 (Q2 V1) (α = .91), and 38 had filled out Q2 V2 (α = .88).

Data analysis. The participants' responses in Van der Kleij and Eggen's (2013) study on the multiple-response items were rescored as separate dichotomous (correct/incorrect) items to obtain maximum information. Furthermore, respondents who had filled out less than 25% in the pre-test were removed from Van der Kleij and Eggen's data set. This resulted in data from 93 respondents, and a questionnaire reliability of α = .95. Subsequently, the results of both questionnaires were analysed using Classical Test Theory (CTT) in TiaPlus (2010). The overall results in terms of proportion correct (P-values) of the interpretation of the original and redesigned report versions were compared. Additional analyses were conducted using IRT. Based on the item parameters from the model, a latent regression model using the computer program SAUL (Verhelst & Verstralen, 2002) was estimated to determine any differences in the abilities of the various user groups—teachers, internal support teachers, or principals.

A differential item functioning (DIF) analysis was also performed, in which the item versions of the redesigned reports were treated as identical to the items belonging to the original reports. Subsequently, we analysed whether particular items functioned differently in the original and redesigned reports.

The results of the items measuring users' perceptions of the redesigned report, compared to those of the original report, were examined using descriptive statistics. The responses were scored from 0 (totally disagree) to 4 (totally agree). Additionally, ANOVA was used to examine any differences amongst the perceptions of respondents from the various user groups, and the relationships with various background variables.

Results. First, the overall results in terms of proportion correct (P-values) were compared. Figure 6.9 shows the P-values of item Version 1 (original reports) versus item Version 2 (redesigned reports) and their 95% CI.

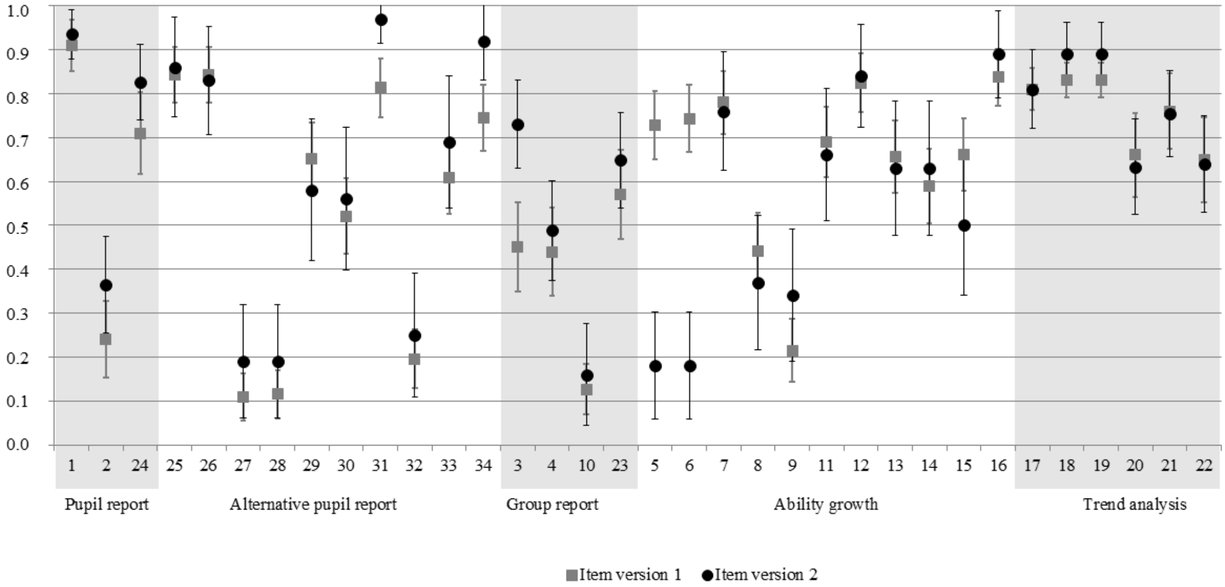


Figure 6.9. P-values of the items in the original and redesigned versions of the reports.

Figure 6.9 shows that three items had significantly larger P-values in the redesigned version of the report, suggesting that they are easier. These items concerned the alternative pupil report and the group report. However, two items were significantly more difficult in the redesigned version; these involved the ability growth report.

Using CTT, the differences in interpretation accuracy amongst various user groups were examined. Table 6.2 shows the results. The results of Q1 and Q2 V2 suggest that internal support teachers and principals interpret the score reports significantly more correctly than teachers do. The results of Q2 V1 are in line with those of Q2 V2, but the interpretation accuracy amongst principals is also higher than those of internal support teachers. However, this result is unreliable, given that only three principals took Q2 V1.

Table 6.2
Average P-values of the User Groups in Q1, Q2 V1, and Q2 V2

Questionnaire version	Q1		Q2 V1		Q2 V2	
User group	*n*	Mean P-value	*N*	Mean P-value	*n*	Mean P-value
Teacher	37	50.32	36	51.65	16	50.82
Internal support teacher	41	68.74	9	68.32	14	67.15
Principal	16	61.03	3	79.41	8	63.44
Total	94	60.29	36	65.06	38	59.35

A 2PL One Parameter Logistic Model (OPLM, 2009; Verhelst & Glas, 1995) was fitted to the data of Q2 (*R1c* = 82.9, *df* = 85, *p* = .54), and the respondents' abilities were estimated. The results of the subsequent analysis in SAUL suggest that internal support teachers and principals have significantly higher abilities to interpret the reports from the Computer Program LOVS than those of teachers (*ES* = 0.9).

An analysis in which the items from Q1 were treated as identical to those in Q2 V1 and Q2 V2—in fact, they were, except for the different image of the report—revealed DIF in three items. This is partially consistent with the results from the CTT analysis. One item belonging to the alternative pupil report was easier, and two items belonging to the ability growth report were harder.

Furthermore, the relationship between various background variables and interpretation ability was examined using SAUL. This computer program allows the structural analysis of a univariate latent variable for complete cases only. Since the information about the various background variables was incomplete for some participants, the analysis had to be carried out separately for the background variables of interest; consequently, the sample sizes differed. The effect sizes (*ES*) reported in the results of the analyses were computed using Equation 1.

$$ES = \frac{\overline{X}_{G1} - \overline{X}_{G2}}{s_{pooled}} \quad (1)$$

There was a significant difference in terms of the number of respondents in each user group who had received training in the use of the Computer Program LOVS in the last five years, *F*(2, 70) = 7.47, *p* = .001. In the teachers' group, only 19% had received training, compared to the internal support teachers (58%) and the majority of the school principals (73%). However, respondents who had received training in the use of the Computer Program LOVS did not significantly perform better in interpreting the redesigned reports (*z* = 0.56, *p* = .809, *ES* = 0.14, *n* = 73).

Also, more years of experience using the Computer Program LOVS did not relate to a higher interpretation ability (0–5, 6–10, >10 years). Taking the lowest category as a reference group, the effects were *z* = 0.91, *p* = .365, *ES* = 0.30, and *z* = 1.24, *p* = .809, *ES* = 0.22 for 6–10 years and >10 years of experience, respectively, with *n* = 73.

The respondents indicated their perceptions of the information generated by the Computer Program LOVS as a little bit useful (*n* = 2), useful (*n* = 38), or very useful (*n* = 33). Participants rating the information as very useful, compared to useful, did not perform significantly better in the interpretation (*z* = 0.71, *p* = .480, *ES* = 0.18). The lowest category was not considered in this case because of few respondents. The ANOVA results show no significant differences amongst respondents in various functions (*F*(2, 70) = 1.67, *p* = .196).

Next, it was investigated whether the respondents' estimates of their own ability in using quantitative test data relate to their ability estimated from the questionnaire responses. Of the 74 respondents, none judged their own ability in using quantitative test data as 'not at all able' or 'not able', 12% deemed themselves as 'a little bit able' (0), 77% as 'able' (1), and 11% as 'very able' (2). Comparing the respondents in categories 0 and 1 revealed a significant relationship with the estimated ability (*z* = 2.27, *p* = .024, *ES* = 0.86). For categories 1 and 2, the relationship with ability was notable (*ES* = 0.94) but not significant (*z* = 1.82, *p* = .069). The estimation of their own ability differed significantly amongst respondents in various functions, *F*(2, 71) = 3.24, *p* = .045. School principals had the highest estimation of their own ability and teachers the lowest.

Subsequently, users' perceptions of the ease in interpreting the original and redesigned versions of the reports were analysed. Table 6.3 shows the results.

Table 6.3

Mean Perceptions of the Respondents Regarding the Redesigned Reports

Report	*n*	Mean (0–4)	*SD*
Group report 1	66	2.38	0.92
Trend analysis	67	2.40	0.78
Group report 2	66	2.50	0.86
Alternative pupil report	30	2.63	0.85
Pupil report 1	71	2.73	0.88
Pupil report 2	65	2.85	0.78
Ability growth	34	2.91	0.75

The mean scores of all the reports are above 2, indicating the majority of the respondents' positive views about the redesigned reports. The respondents were least positive about the group report and trend analysis, and most positive about the pupil report and the ability growth report.

No differences were found amongst the perceptions of the various user groups, except for the report on ability growth, $F(2, 31) = 5.58$, $p = .008$. Here, internal support teachers were significantly more positive than teachers.

6.6.2 Focus Groups

Measurement instruments and procedure. Through 45-minute focus group meetings at two schools, qualitative data were gathered about the users' perceptions of the redesigned reports. The sessions were set up in a group discussion format (Newby, 2010), similar to the previous meetings. The educational adviser fulfilled the role of moderator and was present to answer assessment-specific questions. The first author took notes and clarified the rationales behind the design solutions, when necessary. For each report, the participants were shown the original version, followed by the aspects in need of change, and a display of the redesigned version. Next, the moderator presented the results from the questionnaire, showing the respondents' perceptions of the redesigned reports. Subsequently, for each report, the participants were encouraged to discuss their thoughts on the redesigned reports. They were asked to indicate whether they would like the changes to be implemented, and whether there were further adaptations or refinements needed.

Furthermore, the experts came up with the idea of optionally organising the ability growth report based on clusters of pupils at the same level. The users' ideas regarding offering this option were also probed.

Respondents. The focus group at School 1 consisted of two teachers and two school principals. The focus group at School 2 comprised three teachers, one ICT teacher/coordinator, an internal support teacher, and an adjunct principal.

Data analysis. The participants' comments and recommendations were summarised for each report. Subsequently, these responses were systematically mapped onto the design table (Table 6.1). This analysis served in determining which design solutions were satisfactory, and which ones needed to be adapted or refined.

Results. The participants were enthusiastic about the redesigned reports and agreed that they should be implemented. Particularly valued were the legends in the pupil report and group report, and the level indicators in the ability growth report. The participants provided some useful suggestions for refinement, which mainly related to the tones of the colours used, and the distinctiveness amongst the colours within one report. There was also a positive reaction to the possibility of organising the ability growth report based on clusters of pupils at the same level. The participants reported that this would be particularly valuable for the school's own evaluation purposes.

However, one aspect of the design that related to the score intervals remained doubtful. The researchers expected that integrating the score intervals in the graph, next to reporting the numeric information in the table, would be helpful (inspired by Brown, 2001; Vezzu, VanWinkle, & Zapata-Rivera, 2012; Wainer, 1996). Specifically, the results obtained by Van der Kleij and Eggen (2013) suggest that many users do not understand what the score intervals mean or how they should use this feature, and the information about it is often ignored. The focus group participants claimed that the score intervals might bother or confuse some users. Participants at School 1 did indicate that visualisation of the score intervals was useful, and that it more clearly showed a pupil's learning trajectory. Nevertheless, the results clearly indicate the users' preference for an option not to display the intervals.

6.6.3 Consultation with Key Informants

In the subsequent update of the Computer Program LOVS, a number of changes were implemented (see for an example Figures 6.10 and 6.11). These revisions, along with the remaining intended changes, were evaluated using key informant interviews (Mellenbergh, 2008; Streiner & Norman, 1995).

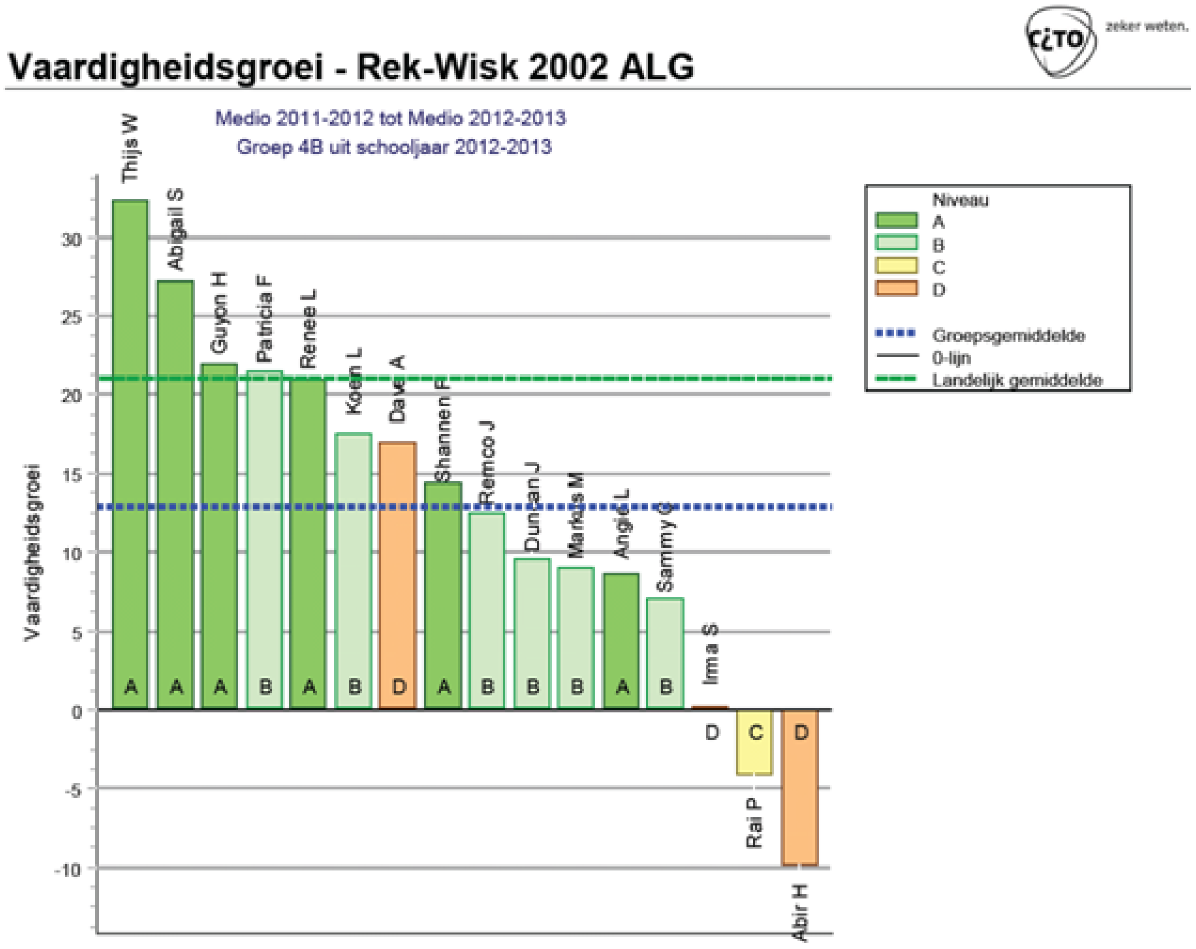


Figure 6.10. Implemented ability growth report.

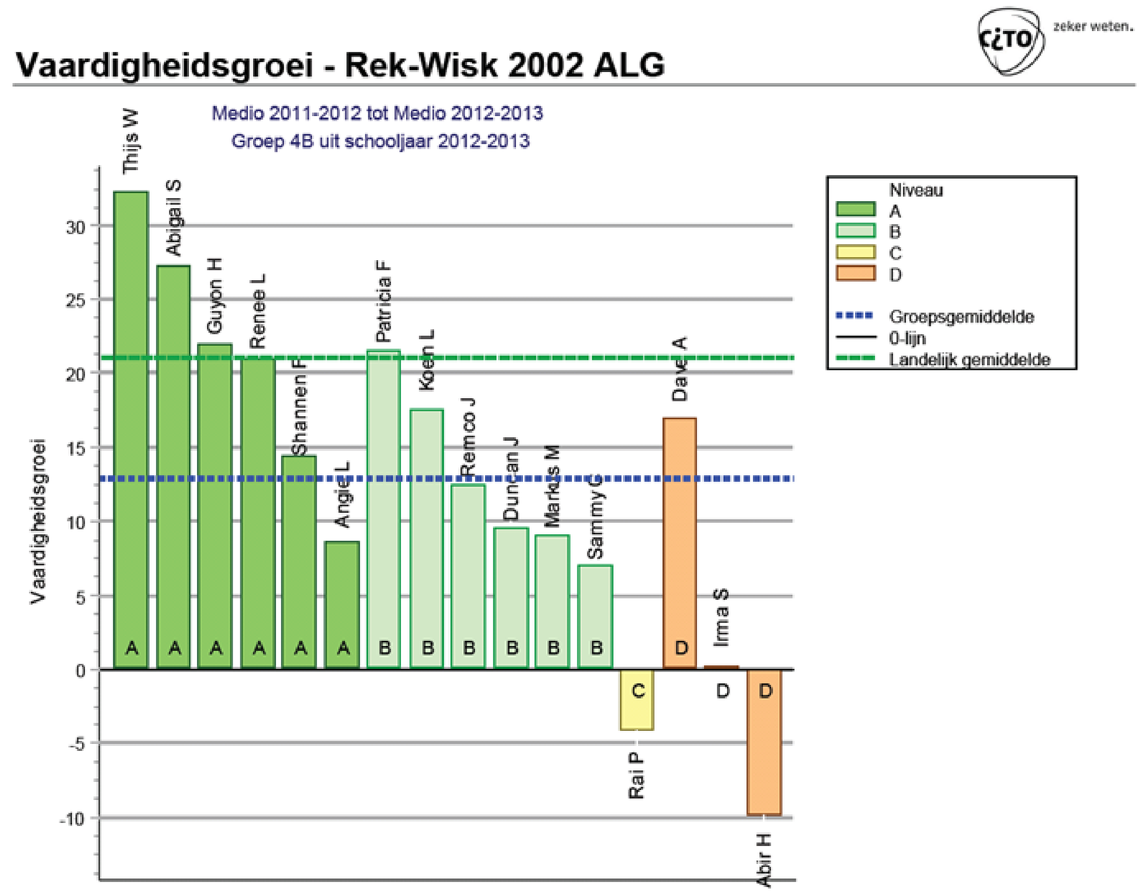


Figure 6.11. Implemented ability growth report, sorted by level of pupils.

Measurement instruments and procedure. Three groups of experienced users were consulted, with a minimum of three years of experience with the Computer Program LOVS. Moreover, the changes were evaluated in consultation with internal support teachers that were non-users of the Computer Program LOVS. These non-users did use the LOVS tests, but used a different computer program for transforming the test scores into score reports. Thus, they were not familiar with the reports of the Computer Program LOVS, but possessed the prerequisite knowledge and skills to interpret these reports. These non-users were consulted to provide an indication regarding the expected effects of the redesigned reports amongst new or inexperienced users.

During face-to-face group interview sessions that lasted between 30 and 45 minutes, in-depth quantitative and qualitative data were gathered. The researcher showed the original and revised reports (both the prototype and the already implemented report) to the participants, along with details of the changes made to the reports. The participants were stimulated to discuss the specific aspects of the redesigned reports with one another and the researcher. The respondents also individually filled out a paper questionnaire, which outlined every change made for each report. For each design solution, they were asked to indicate whether they thought that the particular change made it easier to interpret the report, and to explain their opinion.

Respondents. The experienced users in the consultation stage with the key informants were teachers (*n* = 2), internal support teachers (*n* = 8), and principals (*n* = 5). Three groups of experienced users were consulted (*n* = 5 for each session). The first group of respondents consisted of internal support teachers. In this session, the non-user internal support teachers also participated (*n* = 15). The second group of respondents consisted of teachers (*n* = 2), an internal support teacher (*n* = 1), and principals (*n* = 2) of one school board. The third group of respondents was composed of internal support teachers (*n* = 2) and principals/policymakers (*n* = 3).

Data analysis. The data analysis primarily focused on the data from the group of experienced users. The data from the non-users were used to cross-validate whether the changes would help novice users of the Computer Program LOVS.

For each change, the respondents were asked to answer the question, "Do you think this is an improvement?" These quantitative data were coded (no = 0, yes = 1). The responses to the individual questionnaires were typed and added to the quantitative data set. Additionally, the transcriptions of the audio recordings were used.

First, the quantitative data were analysed. A cut-off score of .80 was used, which means that a greater than.80 agreement rate amongst users would lead to a positive advice for the proposed change. The qualitative data were used to determine whether any minor refinements to the proposed change were needed. For the aspects in which the agreement rate was below .80, it was examined whether and how the change could be best implemented, using the qualitative data from both users and non-users. Furthermore, using the quantitative data, it was evaluated whether there were differences between the opinions of users and non-users for each aspect.

Results. Within the group of users, the mean agreement rate across all reports ranged from .54 to 1 (*M* = .89, *SD* = .32). In the group of non-users, the mean agreement rate ranged from 0.67 to 1 (*M* = .91, *SD* = .28).

In the pupil report, the mean agreement rate was below .80 for only one aspect, i.e., .79 for the changes made to the score intervals. Some users found the display of the score intervals in the graph bothering or confusing, especially for parents. Other users expressed that while they considered this feature useful, the option for hidden score intervals should also be available. Some of the non-users admitted they did not see the added value of the score intervals.

In the report on ability growth, none of the mean scores was below .80. Nevertheless, the qualitative data provided useful advice for refining the design solutions. For example, the labelling of one of the axes appeared confusing, and the users suggested useful alternatives.

The lowest agreement rates were found in the group report, where three of the five aspects were rated below .80. These changes concerned the addition of "comparison all schools" and the black frame around the level of the group, which were supposed to illustrate that the norms for groups differ from those for individual pupils with regard to the level indicators (see Figure 6.8). However, the users indicated that the proposed design solution did not support correct interpretations to a sufficient degree, since the display of the group's ability is a weighted mean, but the accompanying level indicator is not because it comes from a different norm table. The users proposed placing the level of the group elsewhere to avoid confusion, or framing the information regarding it. Furthermore, some of the respondents expressed dissatisfaction with the proposed changes in the use of colours within the level indicators A–E and I–V (.54 mean agreement rate). The users' opinions concerning the use of colours varied widely. Some users preferred a particular colour set because of its non-normative character (e.g., ranging from blue [highest-scoring pupils] to brown [lowest-scoring pupils]). However, others preferred the colours green to red because they provided a natural interpretation and a good signalling function, consistent with Brown's (2001) research findings.

The agreement rates within the report on ability growth (Figures 6.10 and 6.11) were very high (>.93 amongst users). Nevertheless, some useful recommendations for minor revisions were provided, for example, with regard to the clarity of the legend.

For the trend analysis, the mean score amongst users was above .80 for all design solutions. The qualitative data offered additional suggestions for the use of colours.

Overall, the consultation with key informants led to helpful advice for enhancing the design solutions. None of the aspects showed a significant difference between users and non-users.

6.7. Conclusion and Discussion

The results from previous research suggest that users of the Computer Program LOVS struggle in interpreting the reports that feed back results on student learning (Meijer, Ledoux, & Elshof, 2011; Van der Kleij & Eggen, 2013). This study examined whether the interpretability of the reports from this Computer Program LOVS could be improved. Using a design research approach (McKenney & Reeves, 2012), the reports were redesigned in various cycles of design, evaluation, and revision.

A questionnaire was administered to assess the users' interpretations of the redesigned reports. No clear differences were found between the original and the redesigned versions of the reports, in terms of the users' interpretation accuracy. Only three items were significantly easier in the redesigned version, which related to the report on ability growth and group report. The relevant design principles were those of appropriate knowledge, salience, discriminability, and perceptual organisation. Surprisingly, two items were found to be more difficult in the revised version. This result might be explained by a change in the scoring procedures for these items. In the report's original version, only growth in ability was shown, whereas in the revised version, the level indicators were also displayed. Nevertheless, the report's primary purpose still focused on growth in ability. However, when the respondents did not indicate that the redesigned report showed information on the level of pupils, it was scored as incorrect. It must also be mentioned that due to the relatively small sample sizes, the confidence intervals were large, making it improbable for statistical significance to be reached. Additionally, some changes made in the reports were only minor, which makes it unlikely for large effects to be found in terms of interpretation accuracy.

Besides, given the current lack of comparable studies in the literature, it is difficult to establish expectations about interpretation accuracy on the redesigned versions of reports. Also, the reports from the Computer Program LOVS have been in use for years, which makes it challenging to assess user interpretation without it being confounded by prior experience.

Nevertheless, this study's results showed no differences in interpretation accuracy between less experienced and more experienced users. In line with the findings from previous research

(Staman et al., 2013; Van der Kleij & Eggen, 2013), the interpretation accuracy amongst internal support teachers and principals was higher than that of teachers. Although the proportion of teachers who had received training in the use of the Computer Program LOVS was larger than that in the Van der Kleij and Eggen (2013) study, the results suggest that the teachers show the highest need for professional development in assessment literacy.

Although no clear effects were found in terms of users' interpretation accuracy, the results regarding users' perceptions of the redesigned reports in both the focus group meetings and key informant consultations were positive. Users, especially internal support teachers, were particularly positive about the changes made to the report on ability growth. This report was adapted using the principles of salience, discriminability, and perceptual organisation. Furthermore, the principles of relevance and appropriate knowledge seemed relevant to improving interpretability as perceived by users. The meetings resulted in valuable advice for upgrading the design solutions. Eventually, a final set of design solutions was proposed.

In the present study, design principles from the literature were applied throughout the design process and found to be very helpful in directing the latter. Nevertheless, these principles left room for various design solutions and possible variations in graphic design. Therefore, we argue that it is an absolute necessity to involve experts and current/future users in the design process to ensure the validity of the reports.

Although the researchers advocate the careful design of score reports in collaboration with the users, it is evident that well-crafted reports can only partially compensate for the lack of basic knowledge (Hambleton & Slater, 1997). The aspects that initially caused many misinterpretations also posed the most difficulty in achieving a satisfactory design solution. These mainly concerned the more complex issues that required statistical knowledge. An example of a problem that could not be solved by redesigning the reports is the score interval issue. The concept of score intervals seemed complicated to many users. Not surprisingly, only a few actually use them (Van der Kleij & Eggen, 2013). These findings are consistent with previous research results (Hambleton & Slater, 1997; Zenisky & Hambleton, 2012), which indicated that confidence levels are often ignored by users who do not perceive them as valuable. However, the Standards (AERA et al., 1999) do prescribe that confidence levels be reported. Nonetheless, when addressing the principle of appropriate knowledge in this situation, it is clear that not all users are familiar with confidence levels on test scores, not even the more able and experienced users. Thus, the value of reporting score intervals is questionable when they are neither understood nor used in the way the test developer intended (Van der Kleij & Eggen, 2013).

A limitation of this study involves its small sample sizes. However, given the nature of the study, the researchers preferred depth over breadth of information. Thus, it was deemed more useful to receive detailed feedback from a small group of users over several versions of the design solution, than general feedback from a large group. Close collaboration with the experts was maintained throughout the design process to ensure the design's suitability to the entire field of users. However, the present study only made it possible to predict the effectiveness of the redesigned reports to a limited extent. The actual effectiveness in terms of interpretation accuracy will become apparent when the redesigned reports are fully implemented. Furthermore, the design solutions employed in this study will have to be consistently applied to the redesign of the other reports in the Computer Program LOVS. It is also not guaranteed that all the design solutions proposed in this study will be technically feasible in the way the researchers advised. Aside from the report's development process, design and format, and contents, the ancillary materials and dissemination efforts (Zenisky & Hambleton, 2012) should also be aligned with the intended uses of the score reports. The ancillary materials could possibly reinforce support in the report aspects that caused users to struggle.

Although the reports from the Computer Program LOVS had been in use for years, clearly, not all users can interpret them correctly. There might be more pupil-monitoring systems being used whose reports are not understood well and are in need of redesign. Moreover, we propose that the thoughtful design of score reports in collaboration with users be an integral element of the general assessment design.

Future research is needed to clarify the extent of educators' professional development needs in terms of the correct interpretation of score reports and test results in general. Specifically, an accurate data interpretation about student learning is a necessary precondition for successfully implementing DDDM, since it concerns the first step in the evaluative cycle. Nevertheless, research suggests that the subsequent steps, such as making decisions on how to adapt instruction, might even be more challenging (Chahine, 2013; Hattie & Brown, 2008; Heritage, Kim, Vendlinski, & Herman, 2009; Timperley, 2009). These skills have been called "pedagogic data literacy" by Mandinach and Jackson (2012). Several researchers have recently addressed the necessity of pedagogical content knowledge as a precondition for effectively acting on assessment results (Bennett, 2011; Heritage et al., 2009; Timperley, 2009). Although a correct interpretation of assessment results is a prerequisite, it provides no guarantees for their adequate use. Further research in this area is therefore warranted.

Finally, despite the growing body of research on effective score reporting (Zenisky & Hambleton, 2012), little effort has focused on the users' actual interpretations of reports. These types of studies are essential to ensure that the users interpret the test results as the test developer intended, a necessary move towards valid reporting practices.

References

Allalouf, A. (2007). Quality control procedures in the scoring, equating, and reporting of test scores. *Educational Measurement: Issues and Practice, 26,* 36–46. doi:10.1111/j.1745-3992.2007.00087.x

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. doi:10.1080/0969594X.2010.513678

Brown, G. T. L. (2001). *Reporting assessment information to teachers: Report of project asttle outputs design.* asTTle Technical Report #15. Auckland, NZ: University of Auckland. Retrieved from http://e-asttle.tki.org.nz/content/download/1471/5946/version/1/file/15.+Outputs+design+2001.pdf

Chahine, S. (2013, April). *Investigating educators' statistical literacy and score report interpretation.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

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

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

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. doi:10.1076/sesi.11.4.453.3561

Goodman, D., & Hambleton, R. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*, 37–41. doi:10.1207/s15324818ame1702_3

Hambleton, R. K., & Meara, K. (2000). Newspaper coverage of NAEP results, 1990 to 1998. In M. L. Bourque & S. Byrd (Eds.), *Student performance standards on the National Assessment of Educational Progress: Affirmation and improvements* (pp. 132–155). Washington, DC: National Assessment Governing Board.

Hambleton, R. K., & Slater, S. C. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Technical Report 430). Los Angeles, CA: National Center 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. doi:10.2190/ET.36.2.g

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112. doi:10.3102/003465430298487

Hattie, J. (2009). Visibly learning from reports: The validity of score reports. *Online Educational Research Journal*. Retrieved from http://www.oerj.org/View?action=viewPDF&paper=6

Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement: Issues and Practice*, *28*, 24–31. doi:10.1111/j.1745-3992.2009.00151.x

Jaeger, R. M. (2003). *NAEP validity studies: Reporting the results of the National Assessment of Educational Progress.* Working paper 2003–11. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Kosslyn, S. M. (2006). *Graph design for the eye and mind.* New York, NY: Oxford University.

Leeson, H. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing, 6*, 37–41. doi:10.1207/s15327574ijt0601_1

Mandinach, E. B., & Jackson, S. S. (2012). *Transforming teaching and learning through data-driven decision making.* Thousand Oaks, CA: Corwin.

Mayer, R. E. (2001). *Multimedia learning*. Cambridge, UK: Cambridge University Press.

McKenney. S., & Reeves, T. C. (2012). *Conducting educational design research*. London, UK: Routledge.

Mellenbergh, G. J. (2008). Surveys. In H. J. Adèr & G. J. Mellenbergh (Eds.), *Advising on research methods: A consultant's companion* (pp. 183–209). Huizen, the Netherlands: Johannes van Kessel.

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.

Mislevy, R. J. (1998). Implications of market-basket-reporting for achievement level setting. *Applied Measurement in Education, 11*, 49–63. doi:10.1207/s15324818ame1101_3

Monaghan, W. (2006). The facts about subscores. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/ Media/Research/pdf/RD_Connections4.pdf

Newby, P. (2010). *Research methods for education.* Harlow, UK: Longman.

Nitko, A. J., & Brookhart, S. M. (2007). *Educational assessment of students* (5th ed.). Upper Saddle River, NJ: Pearson Education.

One Parameter Logistic Model (Version 2009) [Computer software]. Arnhem, the Netherlands: Cito

Ryan, J. M. (2003). *An analysis of item mapping and test reporting strategies.* Greensboro, NC: SERVE.

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.

Sadler, R. D. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119–144. doi:10.1007/BF00117714

Schildkamp, K., Lai, M. K., & Earl, L. (Eds.). (2013). *Data-based decision making in education: Challenges and opportunities.* Dordrecht, the Netherlands: Springer. doi:10.1007/978-94-007-4816-3

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.

Streiner, D. L., & Norman, G. R. (1995). *Health measurement skills: A practical guide to their development and use* (2nd ed.). Oxford, UK: Oxford University.

Stobart, G. (2008). *Testing times: The uses and abuses of assessment.* Abingdon, England: Routledge.

TiaPlus (Version 2010) [Computer software]. Arnhem, the Netherlands: Cito.

Trout, D. L., & Hyde, E. (2006, April). *Developing score reports for statewide assessments that are valued and used: Feedback from K-12 stakeholders*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

Tufte, E. R. (1990). *Envisioning information.* Cheshire, CT: Graphics Press.

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. doi:10.1080/03055698.2010.482771

Van der Kleij, F. M. & Eggen, T. J. H. M. (2013). Interpretation of the score reports from the Computer Program LOVS by teachers, internal support teachers and principals. *Studies in Educational Evaluation, 39*, 144–152. doi:10.1016/j.stueduc.2013.04.002

Verhaeghe, G. (2011). *School performance feedback systems: Design and implementation issues.* (Doctoral dissertation, University of Gent, Belgium). Retrieved from http://users.ugent.be/~mvalcke/CV/PhD%20Goedele%20Verhaeghe.pdf

Verhelst, N. D., & Glas, C. A. W. (1995). The one-parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 215–237). New York, NY: Springer.

Verhelst, N. D., & Verstralen H. H. F. M. (2002). *Structural analysis of a univariate latent variable (SAUL); theory and a Computer Program*. Arnhem, the Netherlands: Cito. OPD Memorandum 2002-1.

Vezzu, M., VanWinkle, W., & Zapata-Rivera, D. (2012). *Designing and evaluating an interactive score report for students*. Princeton, NJ: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/RM-12-01.pdf

Wainer, H. (1996). Depicting error. *The American Statistician, 50*(2), 101–111. Retrieved from http://www.jstor.org/stable/2684419

Wainer, H. (1997). Improving tabular displays, with NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Statistics, 22*, 1–30. doi:10.3102/10769986022001001

Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement, 36*, 301–335. doi:10.1111/j.1745-3984.1999.tb00559.x

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. doi:10.1111/j.1745-3992.2012.00231.x

Chapter 7. Data-Based Decision Making, Assessment for Learning, and Diagnostic Testing in Formative Assessment[11]

Abstract

Recent research has highlighted the lack of a uniform definition of formative assessment, although its effectiveness is widely acknowledged. This study addresses the theoretical differences and similarities amongst three approaches to formative assessment that are currently most frequently discussed in educational research literature: Data-based decision making (DBDM), assessment for learning (AfL), and diagnostic testing (DT). Furthermore, the differences and similarities in the implementation of each approach were explored. This study shows that although differences exist amongst the theoretical underpinnings of DBDM, AFL, and DT, possibilities are open for implementing an overarching formative assessment and formative evaluation approach. Moreover, the integration of the three assessment approaches can lead to more valid formative decisions. Future research is needed to examine the actual implementation of a mix of these three approaches, with their associated challenges and opportunities.

[11] This chapter has been submitted as Van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., Eggen, T. J. H. M. (submitted). *Data-based decision making, assessment for learning, and diagnostic testing in formative assessment.* Manuscript submitted for publication.

7.1 Introduction

The complex interdependencies amongst learning, teaching, and assessment are increasingly being recognised. Assessment encompasses the use of a broad spectrum of instruments for gathering information about student learning, such as paper-and-pencil tests, projects, or observations (Stobart, 2008). In education, a distinction is made between summative assessment and formative assessment.

Whenever assessment results play a role in making (part of) a decision about the mastery of a defined content domain, it fulfils a summative function, for example, in making a decision regarding selection, classification, certification, or placement (Sanders, 2011). If assessment results are used to steer the learning process, assessment fulfils a formative function. Summative and formative assessments are not mutually exclusive in their purposes; they can coexist as primary and secondary purposes of the same assessment (Bennett, 2011). The effectiveness of formative assessment is widely acknowledged. However, these effectiveness claims are not always well grounded, which is, amongst other things, caused by the lack of a uniform definition of the concept of formative assessment (Bennett, 2011). Formative assessment can be seen as an umbrella term that covers various approaches to assessment that have different underlying learning theories (Briggs, Ruiz-Primo, Furtak, Shepard, & Yin, 2012). The term *approach* captures the underlying principles and intentions that shape particular assessment uses.

Furthermore, it is helpful to make a distinction between formative evaluation and formative assessment (Harlen, 2007; Shepard, 2005). The term *formative evaluation* refers to the use of assessment data to make decisions concerning the quality of education at a higher aggregation level than the level of the learner or the class. Data that are at hand from summative assessment can also be used for formative evaluation (e.g., the use of assessment data for policy development at the school level). *Formative assessment*, on the contrary, only concerns decisions at the levels of the learner and the class to accommodate the pupils' individual educational needs.

This study addresses the theoretical differences and similarities amongst three approaches to formative assessment that are currently most frequently discussed in educational research literature. The main feature that these approaches have in common is that the evidence gathered using assessments is interpreted and subsequently used to change the learning environment in order to meet learners' needs (Wiliam, 2011). However, the way student learning is defined differs within each approach. The first approach is *data-based decision making* (DBDM), which originated in the USA as a direct consequence of the No Child Left Behind (NCLB) Act, which defines improving students' learning outcomes in terms of results and attaining specified targets (Wayman, Spikes, & Volonnino, 2013). Second, *assessment for learning* (AfL), originally introduced by scholars from the UK (Assessment Reform Group [ARG], 1999), is an assessment approach that focuses on the quality of the learning process, rather than merely on students' (final) learning outcomes (Stobart, 2008). Finally, *diagnostic testing* (DT) was initially used to refer students to special education, particularly those diagnosed as unable to participate in mainstream educational settings (Stobart, 2008). In DT, detailed assessment data about a learner's problem solving are collected to explain his or her learning process and learning outcomes (Crisp, 2012; Keeley & Tobey, 2011).

Within some of the approaches, the terminology and definitions are inappropriately used interchangeably; therefore, it is valuable to review and compare the theoretical underpinnings of DBDM, AfL, and DT. For example, literature on DBDM tends to cite literature concerning AfL, but not vice versa (e.g., Swan & Mazur, 2011). Moreover, the discussions in the assessment literature tend to revolve around finding evidence of what works. As Elwood (2006) pointed out, these discussions do not acknowledge the complexity of the use of assessment for learning enhancement, and lead to what she calls "quick fixes" (p. 226). Ignoring the differences in the theoretical underpinnings has led to theoretical ambiguity in the assessment literature, as shown by studies that use the terminology and definitions interchangeably. Bennett (2011) also stressed this ambiguity in the use of definitions: "Definition is important because if we can't clearly define an innovation, we can't meaningfully document its effectiveness" (p. 8). To move the field of educational assessment forward, clarity on the theoretical underpinnings is necessary. More importantly, the relation between these underpinnings and the prescriptions on *why*, *how*, and *when* assessment should be used by learners, teachers, and schools to enhance learning, is needed. Currently, a mix of these approaches is implemented in educational practice. As a result, it is not feasible to study the effectiveness of each approach separately. It is not possible to study which underlying mechanisms contribute to student learning and to what extent. For example, in British Columbia, Canada, DBDM and AfL are the pillars of the programme of the Ministry of Education (2002). However, while DBDM focuses on *what* has to be learned, AfL and DT seem to emphasise *how* students learn what has to be learned (best), and the quality of the learning process (Stobart, 2008). Nevertheless, all three approaches claim the importance of using feedback for learning enhancement, but the procedures advised regarding the provision of feedback differ substantially.

It is important to recognise the fundamental differences amongst these approaches as an initial exploration of what it might mean to blend these approaches in a meaningful way. With this comparative theoretical study, we aim to contribute to a more coherent research agenda within the field of the effectiveness of educational assessment programmes based on these approaches. Note that we do not intend to make any claims about which assessment approach is most effective for improving student learning.

The following questions guide the comparison presented in this study:

1. What are the similarities and differences in the theoretical underpinnings of DBDM, AfL, and DT?
2. What are the consequences of these similarities and differences for implementing DBDM, AfL, and DT in educational practice?

7.1.2 Learning Theories and the Role of Feedback

It is remarkable that most literature about assessment approaches rarely makes explicit the theoretical assumptions about learning (Stobart, 2008). For example, the results of a recent review on formative assessment suggest that the studies which clearly relate formative assessment to a learning theory are scarce (Sluijsmans, Joosten-ten Brinke, & Van der Vleuten, submitted). Implementing a system-wide formative assessment approach requires an alignment of assessment practices, which starts with an understanding of the learning theories behind currently dominant approaches (Elwood, 2006). We considered five learning theories relevant to our comparison of the three assessment approaches, as these are most prominent in the current assessment literature: Neo-behaviourism, cognitivism, meta-cognitivism, social cultural theory, and (social) constructivism (Boekaerts & Simmons, 1995; Stobart, 2008; Thurlings, Vermeulen, Bastiaens, & Stijnen, 2013; Verhofstadt-Denève, Van Geert, & Vyt, 2003).

The formative concept originates from *neo-behaviourism,* as introduced by Bloom, Hastings, and Madaus (1971). Starting from the 1930s, this had been the dominant theory of learning, which has focused on behavioural rather than cognitive mechanisms of learning (Stobart, 2008; Verhofstadt-Denève et al., 2003). Assessment emphasises memorisation of facts, and feedback is intended to reinforce correct recall of these facts (Hattie & Gan, 2011; Narciss, 2008). These facts are seen as independent of the context in which they have been taught (Stobart, 2008).

On the contrary, *cognitivists* such as Piaget focus on changes in cognitive structures rather than in behaviour (Verhofstadt-Denève et al., 2003). Cognitivism highlights information processing and knowledge representation in the memory, rather than learning mechanisms (Shuell, 1986). Because the outcome of learning is still behavioural change, the accompanying assessment and teaching practice are primarily of a retroactive nature, meaning that remediation is used to redirect the learning process and promote learning (Stobart, 2008). Feedback is often intended to correct incorrect responses (Kulhavy & Stock, 1989; Thurlings et al., 2013). However, the characteristics of the learner and the task are taken into account. An expert usually provides the feedback to a passive learner (Evans, 2013).

In *meta-cognitivism*, the emphasis is on learning how to learn and regulating the learning processes by regularly providing feedback (Butler & Winne, 1995). Assessment is aimed at metacognitive knowledge and skills. The feedback message is usually about *how* the learner learns, rather than about *what* the learner learns (Brown, 1987; Stobart, 2008).

Vygotsky (1978) introduced the *social cultural theory* of learning, in which feedback in the form of scaffolding is the most important learning mechanism for acquiring knowledge and skills. Scaffolding is the mechanism in which sociocultural environments facilitate learners' use of knowledge and skills, which they are not yet able to apply on their own. Through social interactions and dialogues between the learner and the teachers, or his or her peers, the learner internalises the knowledge and skills. Vygotsky believed that to promote student learning, assessments should focus on what students are able to learn, rather than what they have learned so far (Verhofstadt-Denève et al., 2003). Although Vygotsky's theory resulted in an international shift in teaching practices, retroactive assessment practices, which focus on remediation, have remained popular (Elwood, 2006; Stobart, 2008). Thus, although learning is seen as a sociocultural interactive activity, assessment remains mostly an individual activity.

In *constructivism,* learning is seen as a cyclic process in which new knowledge and skills are built on prior ones through continuous adaption of the learning environment to the learners' needs (Jonassen, 1991; Stobart, 2008). In *social constructivism*, the learners' active role is emphasised, and teachers are expected to actively engage learners in constructing knowledge and developing skills by frequently providing elaborated feedback. Collaborative learning and solving real-world problems, which use peer feedback as an important learning mechanism, characterise social constructivist learning environments (Lesgold, 2004; Stobart, 2008; Thurlings et al., 2013).

7.2 Theoretical Framework

First, this paper discusses the theoretical underpinnings of each approach in terms of its origin, definition, goals, and relation with the five learning theories. This is followed by a description of the implementation of each approach in terms of aggregation level, assessment methods, and feedback loops.

7.2.1 Theoretical Underpinnings of DBDM

Teachers usually make instructional decisions intuitively (Ingram, Louis, & Schroeder, 2004; Slavin, 2002, 2003). However, educational policies such as No Child Left Behind (NCLB) have caused an increase in accountability requirements, which has stimulated the use of data for informing school practice in the USA (Wayman, Jimerson, & Cho, 2012). Using data to inform decisions in the school is referred to as DBDM (Ledoux, Blok, Boogaard, & Krüger, 2009). Schildkamp and Kuiper (2010) defined DBDM as "systematically analyzing existing data sources within the school, applying outcomes of analyses to innovate teaching, curricula, and school performance, and, implementing (e.g., genuine improvement actions) and evaluating these innovations" (p. 482). The definition of *data* in the context of schools is "information that is systematically collected and organized to represent some aspect of schooling" (Lai & Schildkamp, 2013, p. 10). This definition is broad and includes any relevant information derived from qualitative and quantitative measurements (Lai & Schildkamp, 2013; Wayman et al., 2012).

Data include not only assessment results, but also other data types, such as student background characteristics. Data use can be described as a complex and interpretive process, in which data have to be identified, collected, analysed, and interpreted to become meaningful and useful for actions (Coburn, Toure, & Yamashita, 2009; Coburn & Turner, 2012). The action's impact is evaluated by gathering new data, which creates a feedback loop (Mandinach & Jackson, 2012).

Early initiatives of DBDM were based on neo-behaviourism and cognitivism (Stobart, 2008), which meant that no explicit attention was paid to the sociocultural environment where learning occurred. Previously, DBDM focused on reaching predetermined goals, checking if the goals had been achieved, and adapting the learning environment where needed. This process was mainly transmissive in nature, meaning that educational facilitators (e.g., teachers) were responsible for delivering adequate instruction to learners. In this view, learning is an individual activity, and assessments are used to check on the individual student's ability (Elwood, 2006). As a consequence of this view, assessment methods used for DBDM, such as standardised tests, do not resemble characteristics of every possible learning context in which the learner could have acquired that what is assessed.

However, lately DBDM seems to move more towards social cultural theory and constructivism, which focuses on continuously adapting learning environments to facilitate and optimise learning processes, taking into account learners' needs and individual characteristics. Thus, instead of just acknowledging the context or controlling for it, the emphasis is on the process of data use within a particular context (Coburn & Turner, 2011; Schildkamp, Lai, & Earl, 2013; Supovitz, 2010).

By using data, teachers can set appropriate learning goals, given students' current achievements. Subsequently, teachers can assess and monitor whether students are reaching their goals, and if necessary, adjust their instruction (Bernhardt, 2003; Earl & Katz, 2006). In this way, DBDM is used for formative assessment. The goal of using assessment data in DBDM is almost always improving student achievement (Schildkamp et al., 2013)

Besides, data can be used for formative evaluation by school leaders and teachers for policy development and school improvement planning, teacher development, and monitoring the implementation of the school's goals (Schildkamp et al., 2013; Schildkamp & Kuiper, 2010).

7.2.2 Implementation of DBDM: Aggregation Level, Assessment Methods, and Feedback Loops

Aggregation level. Data collection regarding DBDM takes place at the school, class, and student levels. Data are gathered from different stakeholders. At the student and class levels, assessment results are an important source of information about how learning processes could be improved for both students and teachers. Students need feedback to choose the most suitable learning strategies in order to achieve the intended learning outcomes, while teachers need data to act on the students' current points of struggle and to reflect on their own teaching practices (Young & Kim, 2010). Data can also be used at the school level for school development purposes, for example, to increase aggregated student achievement (Schildkamp & Lai, 2013).

Assessment methods. Different data types can be used for school and instructional development. The data type most often referred to is objective output data from standardised tests, for example, from a student monitoring system. However, these data are less frequently available than those from informal assessment situations, such as homework assignments. Next to these formally gathered data, teachers possess data collected using various standardised assessment methods and (structured) observations from daily practice (Ikemoto & Marsh, 2007).

Access to high-quality data is essential for DBDM because the quality of the decision depends upon the quality of the data used (Coburn & Turner, 2011; Schildkamp & Kuiper, 2010). For the implementation of DBDM, schools need access to multiple sources of high-quality data (especially if the stakes are high) and therefore need a good data use infrastructure (Breiter & Light, 2006; Wayman & Stringfield, 2006).

Feedback loops. The most frequently used kind of feedback in DBDM is feedback based on assessment data. Teachers and other educators have to transform assessment data into meaningful actions for educational improvement. These actions include making changes in practice and providing students with feedback on their learning processes and outcomes (Schildkamp et al., 2013). Often, feedback aims to identify the current achievement level related to the desired achievement level and how to decrease this gap (Hattie & Timperly, 2007). According to Timperly (2009), teachers need to engage in a continuous cycle of inquiry: Based on (assessment) data, identifying what students need to learn and what teachers themselves need to learn, and after taking actions in the form of changes in instruction and/or feedback to students, checking these actions' impacts on the learners. The length of these cycles and feedback loops varies. For example, the feedback loops are relatively long when involving the use of standardised assessments, which are only available once or twice a year. The majority of these loops are retroactive in nature, meaning that based on data, achievement gaps are identified and addressed.

7.2.3 Theoretical Underpinnings of AfL

Assessment for learning (AfL) was originally introduced by UK scholars as a resistance to the emphasis on summative uses of assessments (Stobart, 2008). This approach focuses specifically on the quality of the learning process instead of on its outcomes. Moreover, "it puts the focus on what is being learned and on the quality of classroom interactions and relationships" (Stobart, 2008, p. 145).

The ARG defined AfL as "… the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there" (2002, p. 2). However, this definition was often misinterpreted (Johnson & Burdett, 2010; Klenowski, 2009). For this reason, Klenowski (2009) reported on what she referred to as a "second-generation definition" of AfL, which will be used in this chapter: "part of everyday practice by students, teachers and peers that seeks, reflects upon and responds to information from dialogue, demonstration and observation in ways that enhance ongoing learning" (p. 264).

Hargreaves (2005) concluded that there are two approaches within AfL, a measurement and an inquiry approach. In the measurement approach, AfL is viewed as an activity that includes marking, monitoring, and showing a level. In this view, (quantitative) data are used to formulate feedback and to inform decisions. Assessment is seen as a separate activity from instruction that shows to what degree a predetermined level has been achieved. This approach resembles the definition of DBDM. In the inquiry approach, AfL is a process of discovering, reflecting, understanding, and reviewing. It is focused on the process, and assessments are integrated into the learning process. Qualitative sources of information, such as observations, demonstrations, and conversations, play an important role. In both approaches, feedback is used to steer future learning. However, in the first approach, feedback might be less immediate and feedback loops less frequent. The AfL approach described in this study leans towards the inquiry approach, as described by Klenowski (2009).

In AfL literature, classroom dialogues are stressed as an important learning activity. This idea is theoretically underpinned by meta-cognitivism, social cultural theory, and social constructivism. Learning is seen as a social activity; learning occurs through interaction. Thus, knowledge and skills are believed to depend on the context, and to exist in the relationships amongst the individuals involved in that context. As a result, assessment should not be seen as an individual activity either (Elwood, 2006).

The AfL approach is aimed at the quality of the learning process instead of its outcomes (e.g., a grade). This goal stimulates a learning-oriented rather than an outcome-oriented classroom culture (Stobart, 2008). AfL makes it possible to anticipate weaker points in the current learning process and identify further steps to take for improvement (ARG, 1999). Students have a central role in the learning process; as a result, they actively participate in the evaluation of their own learning (Elwood & Klenowski, 2002). Furthermore, AfL aims to increase learner autonomy, motivation, and reflection, by facilitating an inquiry-oriented and interactive classroom climate (Klenowski, 2009).

7.2.4 Implementation of AfL: Aggregation Level, Assessment Methods, and Feedback Loops

Aggregation level. The AfL approach takes place within the classroom; it concerns decisions about the entire class or individual students. The information used to make decisions is gathered from students.

Assessment methods. The data used to inform decisions can come from various assessment sources, such as paper-and-pencil tests, dialogues, practical demonstrations of learning, portfolios, peer assessment, or self-assessment (Gipps, 1994). Hence, the evidence gathered about the learning process of the learners can be both qualitative and quantitative in nature. These assessment events can be planned as well as unplanned, and formal and informal. Continuous interactions between learners and the teacher characterise the process.

The quality of the assessment process depends largely on the degree of the teacher's capability to obtain usable data about student learning, make inferences about student learning, and translate this information into instructional decisions (Bennett, 2011). Thus, the assessment quality depends on the degree to which assessment results provide actionable information for formative purposes over the short term, which is a low-stakes type of use. Nevertheless, teachers' inferences are likely to be biased to some extent (Bennett, 2011); therefore, standardised assessments can be used once in a while to check on students' learning outcomes in terms of overall curriculum goals and standards.

Feedback loops. AfL takes place in everyday practice; continuous dialogues and feedback loops characterise the process, in which (immediate) feedback is used to direct further learning (Stobart, 2008). Since assessments are integrated into the learning process, assessment opportunities are plentiful, and feedback loops are usually short. Moreover, students are stimulated to assess themselves and their peers, which, amongst other things, stimulates students' understanding of what and why they are learning (Elwood & Klenowski, 2002). Based on the evidence gathered, continuous adaptation takes place to meet learners' needs. Thus, the majority of the feedback loops are interactive in nature, but retroactive or proactive loops also occur.

7.2.5 Theoretical Underpinnings of DT

Making diagnoses originates from the field of physical and mental healthcare, which aims to diagnose a disease or disorder and to advise on the treatment (Kievit, Tak, & Bosch, 2002). In education, diagnostic testing (DT) was initially used for identifying students who were unable to participate in mainstream education due to their special educational needs (Stobart, 2008). The DT approach still serves the above purpose, but is also acknowledged for its possibilities to diagnose the educational needs of *all* learners (Wiliam, 2011).

Consensus regarding the definition of DT in educational contexts has yet to be reached. In some of the literature, DT is used as a synonym for formative assessment (e.g., Black & Wiliam, 1998; Turner, VanderHeide, & Fynewever, 2011). However, similar to DBDM, DT can be used for both formative assessment and summative assessment. The assumption in DT is that *how* a task is solved is indicative of the developmental stage of the learner. Collecting data about the procedural steps the learner takes during an assessment can identify the learner's (inadequate) reasoning styles, and skipped or wrongly executed procedural steps caused by misconceptions and prior knowledge, amongst other things (Crisp, 2012; Keeley & Tobey, 2011).

In DT, principles from cognitive psychology, subject pedagogy, and learning theories are combined to draw inferences about student learning based on a student's task response patterns. Using cognitive psychology makes it evident that DT is based on principles from cognitivism (Leighton & Gierl, 2007a; 2007b). Furthermore, Stobart described diagnosing student learning needs as "… [identifying] how much progress can be made with adult help…" (2008, p. 55) (i.e., zone of proximal development; Vygotsky, 1978). The fine-grained process data obtained with DT are particularly useful for creating scaffolds that meet the learner's needs. In this way, DT is related to Vygotsky's social cultural learning theory, where assessment focuses on identifying the learner's strengths and weaknesses.

The aim of DT is to identify the learner's developmental stages by obtaining action-oriented, fine-grained assessment data, also referred to as process data (Rupp, Gushta, Mislevy, & Shaffer, 2010). By using cognitive theories, process data can be interpreted and used to identify misconceptions and knowledge associated with the learner's developmental stage. The intended small-grain size of the measurements in DT, compared to regular assessments, makes it exceptionally useful for formative purposes.

7.2.6 Implementation of DT: Aggregation Level, Assessment Methods, and Feedback Loops

Aggregation level. DT concerns the assessment of the educational needs of individual students. Because of the nature of the instruments used in DT, data should not be aggregated to levels higher than the individual level (Rupp et al., 2010). Furthermore, DT is not meant for comparing students to one another, but for promoting the learning and developmental process of individual students.

Assessment methods. In order to make inferences about the problem-solving process during an assessment, the assessment tasks should be designed to make possible valid inferences about how the student's task behaviour relates back to his or her thinking. This inferential chain stems from the empirical knowledge available from information processing theories, cognitive psychology, and learning trajectories (Daro, Mosher, & Corcoran, 2011; Leighton &

Gierl, 2007a; Verhofstadt-Denève et al., 2003). Based on theoretical assumptions and empirical research, the items in an assessment have certain characteristics that are assumed to elicit a response behaviour related to the learner's developmental stage (Leighton & Gierl, 2007a).

The degree to which the assessment results are indicative of the developmental stage of a student is crucial to the quality of the assessment methods used in DT. Although including more items with the same characteristics in the assessment will increase the certainty of the inferences about related misconceptions, it will also make the assessment process less efficient (Rupp et al., 2010). For example, if the aim is to identify an arithmetic misconception, and a student makes an associated error on one item, it is possible that this error is caused by something else than that particular misconception. However, when the student consistently shows the same error on several items with similar characteristics, the inference about the misconception becomes stronger. Nevertheless, choosing details over certainty, in terms of test accuracy, is not problematic with short feedback loops, because the latter provides the opportunity to redirect the decisions made. In this case, the stakes in terms of possible negative consequences for the learner are relatively small (Rupp et al., 2010).

Moreover, to cope with this trade-off between grain size and certainty about inferences, assessment developers in DT often consider the design of (computerised) adaptive tests, meaning that the selection of the next item depends on the student's response to the previous item (Eggen, 2004). Adaptivity offers the possibility to make the assessment process more efficient; items can be chosen based on their content and difficulty, for example, to diagnose a student's strategy choice. Sometimes these types of assessments are referred to as dynamic assessments, which are usually embedded in a computerised adaptive learning environment. This means that when a student cannot solve a task, he or she will receive a minimally intrusive hint. In this way, the materials are used for both assessment and learning, by providing diagnostic information about a student's learning needs and item-based feedback (Stevenson, Hickendorff, Resing, Heiser, & de Boeck, 2013).

Feedback loops. Although DT has the potential to be used for retroactive, proactive, or interactive formative assessment, it is primarily used retroactively (Crisp, 2012; Stobart, 2008). In dynamic assessments, DT is used interactively; learning and assessment are integrated. When DT focuses on the assessment of prior knowledge to plan instruction, it is used proactively. Finally, when DT is used to identify, for example, misconceptions or buggy problem-solving strategies, feedback is used for remediation, resulting in a retroactive feedback loop. Short feedback loops in DT are preferred because the learner's thinking and use of problem-solving strategies are highly likely to change over time. However, delayed feedback could still be effective when the change in the learner's thinking and the development of new strategies cover longer periods of time. Thus, the length of feedback loops should match the student's learning curve for the subject matter that is the assessment's objective. A mismatch between the two might result in negative consequences, hindering the optimisation of the learning process.

7.3 Comparison of the Three Approaches

This section addresses the theoretical differences and similarities amongst the three approaches. Furthermore, the differences and similarities in the implementation of each approach are explored.

7.3.1 Theoretical Underpinnings of DBDM, Afl, and DT

To answer our first question (What are the similarities and differences in the theoretical underpinnings of DBDM, AfL, and DT?), we compared the underlying learning theories of these approaches and their goals (Table 7.1).

Table 7.1

Comparison of DBDM, AFL, and DT regarding the Theoretical Underpinnings

	Approach		
Theoretical aspect	DBDM	AfL	DT
Learning theories	• Neo-behaviourism • Social cultural theory • (Social) constructivism	• Meta-cognitivism • Social cultural theory • Social constructivism	• Cognitivism • Social cultural theory
Goals	• Improve the quality of education and the quality of instruction by using data to monitor and steer practices to reach intended goals (e.g., increased student achievement).	• Improve the quality of the learning process by engaging learners to evaluate and reflect on their own learning and steering the learning process through continuous feedback.	• Collect fine-grained data about a student's zone of proximal development, prior knowledge, and reasoning styles that can inform decisions on adapting the learning environment to the learner's needs.

Table 7.1 shows that DBDM, AfL, and DT are underpinned by elements from different learning theories. Consequently, the goals of the three approaches differ substantially; each approach aims to promote learning through different mechanisms, which results in different expectations of the roles of teachers, students, and other actors in the learning, assessment, and feedback process. These expectations sometimes contradict each other; for example, in traditional views on DBDM, the responsibility for the assessment process is primarily in the teacher's hands, whereas in AfL, the teacher and students share this responsibility, e.g., in the form of self- and peer assessment (Stobart, 2008; Wiliam, 2011).

However, recent literature on DBDM shows a shift towards shared teacher-student responsibility for assessment (Schildkamp et al., 2013).

We observed that all three approaches converge, following a view of learning based on the principles of meta-cognitivism, social cultural theory, and (social) constructivism. Central to these learning theories is the acknowledgement that most knowledge and skills are inextricably connected to the context in which they are taught. Nevertheless, as described in the next section, this convergence does not affect the implementation of each assessment approach to the same degree.

7.3.2 Implementation of DBDM, AfL, and DT

To answer our second question (What are the consequences of these similarities and differences for implementing DBDM, AfL, and DT in educational practice?), we compared the aggregation levels, assessment methods, and feedback loops of the three approaches (Table 7.2). Figure 7.1 shows the overlapping levels of the decisions in the three approaches. In DBDM, data are aggregated at the school level to make decisions with regard to improving the school's quality (formative evaluation), as well as to judge the latter (summative evaluation). Additionally, data are used at the class and student levels to adjust instruction to meet the students' needs (formative assessment). The latter overlaps with AfL. DT solely focuses on assessment and instructional decisions at the student level. Because the three approaches aim to promote learning at different aggregation levels, they can complement each other.

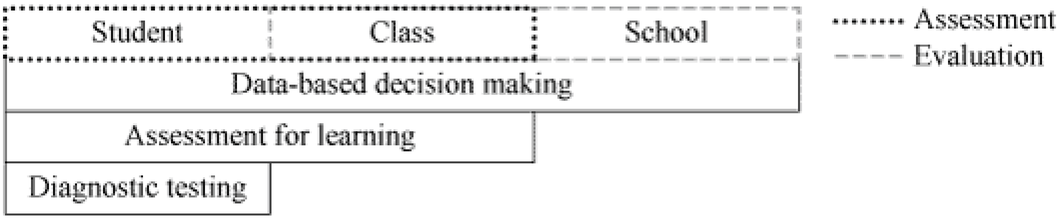


Figure 7.1. Overlapping levels of the decisions in the three approaches.

The diversion in the goals of DBDM, AfL, and DT is associated with a large variety of the use of assessment methods. For example, AfL uses classroom conversations, while DBDM and DT often employ standardised tests. The different choices in the use of assessment methods are primarily associated with the nature of the data, and the purposes and stakes regarding the use of these data. In DBDM, most data are quantitative in nature; especially at the school level, high-quality data are needed as the stakes are often higher. In contrast, most data are qualitative in AfL because they mainly aim to provide immediate information, which informs decisions on how to direct learning processes. These are low-stakes decisions; if the adaptations in the learning environment do not produce the intended effects, this will become quickly clear from subsequent assessments, whereupon the adaptation strategy can be changed. Thus, the adaptation process is flexible. In DT, fine-grained, quantitative data are usually gathered and translated into qualitative statements on which teachers can take immediate actions. Although DT uses quantitative data similar to DBDM, the quality requirements are different from those of DBDM. The stakes associated with DT decisions are usually lower because the decisions concern individuals, and the feedback loops are shorter.

With respect to feedback mechanisms, we found the use of feedback loops in all three approaches. However, because the approaches aim at formative assessment and formative evaluation at different levels, these feedback loops also take place at various levels and frequencies. In DBDM, the retroactive feedback loops that occur at the school level are spread out over time. In AfL, continuous dialogues and feedback loops are essential, which results in short, frequently interactive, and sometimes retroactive or proactive feedback loops. Regarding DT, the length of feedback loops should match the student's learning curve for the subject matter that is the assessment's objective.

Table 7.2
Comparison of the Implementation of DBDM, AfL, and DT

	Approach		
Implementation characteristic	DBDM	AfL	DT
Level	• School • Class • Student	• Class • Student	• Student
Assessment methods	• Standardised assessments • Formal classroom assessments • Structured classroom observations	• Informal classroom dialogues • Formal classroom assessments • Practical demonstrations • Portfolios • Peer assessments • Self-assessments	• (Adaptive) tests with items that elicit detailed information about a student's reasoning
Feedback loops	• Immediate and delayed feedback • Retroactive	• Immediate feedback • Interactive, sometimes retroactive or proactive	• Immediate and delayed feedback • Mostly retroactive, potentially proactive or interactive

7.4 Discussion

The DBDM, AfL, and DT approaches are all seen as powerful ways to support and enhance student learning. Educational practice implements a mix of these approaches. The differences amongst the implementation of assessment approaches stem from differences in their theoretical underpinnings (Stobart, 2008). This study compared the similarities and differences in the theoretical bases of DBDM, AfL, and DT. The differences and similarities in implementing each approach were also explored.

Our comparison suggests that the original theoretical underpinnings of the approaches differ in their definitions of learning. Nevertheless, all approaches increasingly recognise that the assessment's focus should be both on the learning process and on learning outcomes. Various assessment methods that are underpinned by different learning theories are needed to fully grasp the complexity of learning at all levels. If one wants to use assessments or evaluations formatively, one should acknowledge which learning mechanisms are applicable for decision making at the school, class, or student level. Integrating the three assessment approaches can lead to more valid formative decisions. Decisions are no longer based on a single data type at one aggregation level, but on multiple data sources gathered from multiple perspectives at different aggregation levels. Integrating the assessment approaches will enable schools to capture various aspects of their curriculum and the different learning activities of their students. Consequently, school staff will be able to continuously provide feedback at the school, class, and individual levels, to guide and enhance student learning.

To blend the approaches, different feedback loops should be simultaneously active on each level in schools. At the school level, DBDM can be used, for example, to monitor set learning goals, to group students differently to enhance learning, and to improve the quality of education. Moreover, DBDM can be applied to monitor student achievement goals at the class level. Similarly, DBDM can be employed to monitor individual progress. The DBDM approach is often connected to the use of standardised external assessments; therefore, feedback loops usually extend over a longer period of time. The AfL approach can be used at the class and individual levels to improve the quality of the learning process by engaging learners to evaluate and reflect on their own learning, and steering the learning process through continuous feedback. Finally, DT can be employed at the individual level to collect fine-grained data about a student's zone of proximal development, prior knowledge, and reasoning styles that can inform decisions on adapting the learning environment to the learner's needs. Feedback loops occur irregularly; the frequency of DT and subsequent feedback depends on the learner's needs. Thus, at different points in the education process, retroactive, interactive, or proactive feedback loops can be used to optimise students' learning processes.

Accountability is a complicating factor in blending the three approaches. Stakeholders (e.g., the government or parents) hold schools, teachers, and students responsible for meeting certain standards. Accountability decisions are based on summative assessment and summative evaluation. Given the stakes of these decisions, using objective, reliable, and valid data is desirable (Harlen, 2010). Data gathered for formative purposes, whether for DBDM, AfL, or DT, should not be used for accountability purposes for the following reasons. First, these data often do not meet the requirements needed for making accurate high-stakes judgements. Second, high-accountability pressure often results in a narrow set of goals and consequently, a limited use of assessment methods. It will be harder or even impossible for teachers to implement DBDM, AfL, and DT because summative uses of assessments tend to overshadow their formative uses. For example, high-accountability pressure may result in strategic practices, such as teaching to the test and focusing all efforts on the bubble kids, who are children on the threshold of passing the test (Booher-Jennings, 2005; Diamond & Spillane, 2004).

This study shows that although differences exist amongst the theoretical underpinnings of DBDM, AfL, and DT, possibilities are open for implementing an overarching formative assessment and formative evaluation approach. We initially explored what it might mean to blend these approaches. Future research is needed to examine the actual implementation of a mix of these three approaches, with their associated challenges and opportunities.

References

Assessment Reform Group. (1999). *Assessment for learning: Beyond the black box.* Retrieved from http://assessmentreformgroep.files.wordpress.com/2012/01/beyond_blackbox.pdf

Assessment Reform Group. (2002). *Assessment is for learning: 10 principles. Research-based principles to guide classroom practice.* Retrieved from http://assessmentreformgroup.files.wordpress.com/2012/01/10principles_english.pdf

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

Bernhardt, V. L. (2003). Using data to improve student achievement. *Educational Leadership, 60*(5), 26–30. Retrieved from http://www.ascd.org/publications/educational-leadership/feb03/vol60/num05/No-Schools-Left-Behind.aspx

Black, P., & Wiliam, D. (1998). Inside the black box. Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139–148. Retrieved from http://blog.discoveryeducation.com/assessment/files/2009/02/blackbox_article.pdf

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning.* New York, NY: McGraw-Hill.

Boekaerts, M., & Simons, P. R. J. (1995). *Leren en instructie: Psychologie van de leerling en het leerproces* [Learning and instruction: Psychology of the student and the learning process]. Assen, the Netherlands: Van Gorcum.

Booher-Jennings, J. (2005). Below the bubble: "educational triage" and the Texas accountability system. *American Educational Research Journal, 42*, 231–268. doi:10.3102/00028312042002231

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. Retrieved from http://www.ifets.info/journals/9_3/18.pdf

Briggs, D. C., Ruiz-Primo, M. A., Furtak, E., Shepard, L., & Yin, Y. (2012). Meta-analytic methodology and inferences about the efficacy of formative assessment. *Educational Measurement: Issues and Practice*, *31*, 13–17. doi:10.1111/j.1745-3992.2012.00251.x/full

Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. E. Weinart & R. H. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 65–116). Hillsdale, NJ: Lawrence Erlbaum.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*, 245–281. doi:10.3102/00346543065003245

Coburn, C. E., Toure, J., & Yamashita, M. (2009). Evidence, interpretation, and persuasion: Instructional decision making in the district central office. *Teachers College Record, 111*(4), 1115–1161. Retrieved from http://www.tcrecord.org/Content.asp?ContentId=15232

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement, 9*, 173–206. doi:10.1080/15366367.2011.626729

Coburn, C. E., & Turner, E. O. (2012). The practice of data use: An introduction. *American Journal of Education, 118*, 99–111. doi:10.1086/663272

Crisp, G. T. (2012). Integrative assessment: Reframing assessment practice for current and future learning. *Assessment & Evaluation in Higher Education*, *37*, 33–43. doi:10.1080/02602938.2010.494234

Daro, P., Mosher, F. A., & Corcoran, T. (2011). *Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction* (CPRE Research Report #RR-68). Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania Graduate School of Education.

Diamond, J. B., & Spillane, J. P. (2004). High-stakes accountability in urban elementary schools: Challenging or reproducing inequality. *Teachers College Record, 106*(6), 1145–1176. Retrieved from http://www.ipr.northwestern.edu/publications/docs/workingpapers/2002/IPR-WP-02-22.pdf

Earl, L. M., & Katz, S. (2006). *Leading schools in a data-rich world: Harnessing data for school improvement.* Thousand Oaks, CA: Corwin.

Eggen, T. J. H. M. (2004). *Contributions to the theory and practice of computerized adaptive testing.* (Doctoral dissertation, University of Twente, Enschede, the Netherlands). Retrieved from http://www.cito.nl/~/media/cito_nl/Files/Onderzoek%20en%20wetenschap/cito_dissertatie_theo_eggen.ashx

Elwood, J. (2006). Gender issues in testing and assessment. In C. Skelton, B. Francis, & L. Smulyan (Eds.), *Handbook of gender and education* (pp. 262–278). Thousand Oaks, CA: Sage.

Elwood, J., & Klenowski, V. (2002). Creating communities of shared practice: The challenges of assessment use in learning and teaching. *Assessment & Evaluation in Higher Education, 27,* 243–256. doi:10.1080/0260293022013860

Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research*, *83*, 70–120. doi:10.3102/0034654312474350

Gipps, C. (1994). *Beyond testing: Towards a theory of educational assessment.* London, UK: Falmer.

Hargreaves, E. (2005). Assessment for learning? Thinking outside the (black) box. *Cambridge Journal of Education, 35,* 213–224. doi:10.1080/03057640500146880

Harlen, W. (2007). *The quality of learning: Assessment alternatives for primary education*. *Interim Reports*. Cambridge, UK: University of Cambridge. Retrieved from http://gtcni.openrepository.com/gtcni/bitstream/2428/29272/2/Primary_Review_Harlen_3-4_briefing_Quality_of_learning_-_Assessment_alternatives_071102.pdf

Harlen, W. (2010). What is quality teacher assessment? In J. Gardner, W. Harlen, L. Hayward, & G. Stobart (Eds.), *Developing teacher assessment* (pp. 29–52). Maidenhead, UK: Open University.

Hattie, J., & Gan, M. (2011). Instruction based on feedback. In P. Alexander & R. E. Mayer (Eds.), *Handbook of research on learning and instruction* (pp. 249–271). New York, NY: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112. doi:10.3102/003465430298487

Ikemoto, G. S., & Marsh, J. A. (2007). Cutting through the data-driven mantra: Different conceptions of data-driven decision making. In P. A. Moss (Ed.), *Evidence and decision making* (pp. 105–131). Malden, MA: Wiley-Blackwell. Retrieved from http://www.rand.org/content/dam/rand/pubs/reprints/2009/RAND_RP1372.pdf

Ingram, D., Louis, S. K., & Schroeder, R. G. (2004). Accountability policies and teacher decision making: Barriers to the use of data to improve practice. *Teachers College Record, 106*(6), 1258–1287. doi:10.1111/j.1467-9620.2004.00379.x

Johnson, M., & Burdett, N. (2010). Intention, interpretation and implementation: Some paradoxes of assessment for learning across educational contexts. *Research in Comparative and International Education, 5*, 122–130. doi:10.2304/rcie.2010.5.2.122

Jonassen, D. H. (1991). Evaluating constructivist learning. *Educational Technology, 31*(9), 28–33.

Keeley, P., & Tobey, C. R. (2011). *Mathematics formative assessment.* Thousand Oaks, CA: Corwin.

Kievit, Th., Tak, J. A., & Bosch, J. D. (Eds.). (2002). *Handboek psychodiagnostiek voor de hulpverlening aan kinderen* (6th ed.) [Handbook psychodiagnostics in healthcare for children]. Utrecht, the Netherlands: De Tijdstroom.

Klenowski, V. (2009). Assessment for learning revisited: An Asia-Pacific perspective. *Assessment in Education: Principles, Policy & Practice, 16,* 263–268. doi:10.1080/09695940903319646

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, the Netherlands: Springer. doi:10.1007/978-94-007-4816-3

Ledoux, G., Blok, H., Boogaard, M., & Krüger, M. (2009). *Opbrengstgericht werken. Over waarde van meetgestuurd onderwijs* [Data-driven decision making. About the value of measurement oriented education]. SCO-Rapport 812. Amsterdam, the Netherlands: SCO-Kohnstamm Instituut. Retrieved from http://dare.uva.nl/document/170475

Leighton, J. P., & Gierl, M. J. (2007a). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, *26*, 3–16. doi:10.1111/j.1745-3992.2007.00090.x

Leighton, J. P., & Gierl, M. J. (Eds.). (2007b). *Cognitive diagnostic assessment for education. Theory and applications.* New York, NY: Cambridge University.

Lesgold, A. (2004). Contextual requirements for constructivist learning. *International Journal of Educational Research, 41,* 495–502. doi:10.1016/j.ijer.2005.08.014

Mandinach, E. B., & Jackson, S. S. (2012). *Transforming teaching and learning through data-driven decision making.* Thousand Oaks, CA: Corwin.

Ministry of Education, British Columbia, Canada. (2002). *Accountability framework.* Retrieved from http://www.bced.gov.bc.ca/policy/policies/accountability_framework.htm

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merril, J. J. G. van Merriënboer, and M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 125–144). Mahwah, NJ: Lawrence Erlbaum Associates.

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *The Journal of Technology, Learning, and Assessment, 8*(4). Retrieved from http://napoleon.bc.edu/ojs/index.php/jtla/article/viewFile/1623/1467

Sanders, P. (2011). Het doel van toetsen [The purpose of testing]. In P. Sanders (Ed.), *Toetsen op school* [Testing at school] (pp. 9–20). Arnhem, the Netherlands: Cito. Retrieved from http://www.cito.nl/~/media/cito_nl/Files/Onderzoek%20en%20wetenschap/cito_toetsen_op_school.ashx

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. doi:10.1016/j.tate.2009.06.007

Schildkamp, K., Lai, M. K., & Earl, L. (Eds.). (2013). *Data-based decision making in education: Challenges and opportunities.* Dordrecht, the Netherlands: Springer. doi:10.1007/978-94-007-4816-3

Shepard, L. A. (2005, October). *Formative assessment: Caveat emptor*. Paper presented at the ETS Invitational Conference, The Future of Assessment: Shaping Teaching and Learning, New York, NY.

Shuell, T. (1986). Cognitive conceptions of learning. *Review of Educational Research, 56*, 411–436. doi:10.3102/00346543056004411

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 21*, 15–21. doi:10.3102/0013189X031007015

Slavin, R. E. (2003). A reader's guide to scientifically based research. *Educational Leadership, 60*(5), 12–16. Retrieved from http://www.ascd.org/publications/educational-leadership/feb03/vol60/num05/A-Reader's-Guide-to-Scientifically-Based-Research.aspx

Sluijsmans, D., Joosten-ten Brinke, D., & Van der Vleuten, C. (submitted). *Toetsen met leerwaarde: Een reviewstudie naar de effectieve kenmerken van formatief toetsen* [Testing with learning value: A review study on the effective characteristics of formative assessment]. Manuscript submitted for publication.

Stevenson, C. E., Hickendorff, M., Resing, W. C. M., Heiser, W. J., & de Boeck, P. A. L. (2013). Explanatory item response modelling of children's change on a dynamic test of analogical reasoning. *Intelligence, 41*, 157–168. doi:10.1016/j.intell.2013.01.003

Stobart, G. (2008). *Testing times: The uses and abuses of assessment.* Abingdon, England: Routledge.

Supovitz, J. (2010). Knowledge-based organizational learning for instructional improvement. In A. Hargreaves, A. Lieberman, M. Fullan, & D. Hopkins (Eds.), *Second international handbook of educational change* (pp. 707–723). New York, NY: Springer. doi:10.1007/978-90-481-2660-6

Swan, G., & Mazur, J. (2011). Examining data driven decision making via formative assessment: A confluence of technology, data interpretation heuristics and curricular policy. *Contemporary Issues in Technology and Teacher Education, 11*(2), 205–222. Retrieved from http://www.editlib.org/p/36021

Thurlings, M., Vermeulen, M., Bastiaens, T., & Stijnen, S. (2013). Understanding feedback: A learning theory perspective. *Educational Research Review*, *9*, 1–15. doi:10.1016/j.edurev.2012.11.004

Timperley, H. (2009). Using assessment data for improving teaching practice. *Australian College of Educators, 8*(3), 21–27. Retrieved fromhttp://oksowhat.wikispaces.com/file /view/Using+assessment+data+Helen+Timperley.pdf

Turner, M., VanderHeide, K., & Fynewever, H. (2011). Motivations for and barriers to the implementation of diagnostic assessment practices – a case study. *Chemistry Education Research and Practice, 12,* 142–157. doi:10.1039/C1RP90019F

Verhofstadt-Denève, L., Van Geert, P., & Vyt, A. (2003). *Handboek ontwikkelingspsychologie. Grondslagen en theorieën* [Handbook developmental psychology. Principles and theories]. Houten, the Netherlands: Bohn Stafleu Van Loghum.

Vygotsky, L. S. (1978). *Mind in society.* London, England: Harvard University Press.

Wayman, J. C., Jimerson, J. B., & Cho, V. (2012). Organizational considerations in establishing the data-informed district, school effectiveness and school improvement. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice, 23*, 159–178. doi:10.1080/09243453.2011.652124

Wayman, J. C., Spikes, D. D., & Volonnino, M. (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). doi:10.1007/978-90-481-2660-6

Wayman, J. C., & Stringfield, S. (2006). Data use for school improvement: School practices and research perspectives. *American Journal of Education*, *112*, 463–468. doi:10.1086/505055

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*, 3–14. doi:10.106/j.stueduc.2011.03.001

Young, V. M., & Kim, D. H. (2010). Using assessments for instructional improvement: A literature review. *Educational Policy Analysis Archives, 18*(19). Retrieved from http://epaa.asu.edu/ojs/article/view/809

Epilogue

Formative assessment concerns any assessment that provides feedback that is intended to support learning and can be used by both teachers and students. Computers could offer a solution to overcoming the obstacles encountered when implementing formative assessment. The focus of this dissertation is the extent to which the computer could support formative assessment practices by investigating three areas:

1. Item-based feedback provided to students through a computer (Chapters 2, 3, and 4)
2. Feedback provided through a computer to educators based on students' assessment results (Chapters 5 and 6)
3) A comparison of three approaches to formative assessment: Data-based decision making (DBDM), assessment for learning (AfL), and diagnostic testing (DT) (Chapter 7)

I would like to emphasise the huge potential role of computers in supporting formative assessment. Indeed, many opportunities are yet to be explored. However, this tool is useless when the process surrounding it is flawed. Specifically, it is essential that high-quality assessment instruments be used for the purpose they intend to serve. However, the persons using such instruments also need to be skilled in many ways in order for computer-based formative assessment to fulfil its potential. The mere availability of high-quality feedback does not guarantee its effectiveness. Following the structure of the three areas covered by this dissertation, I provide a succinct synthesis of the theoretical and practical findings of this research and their implications for future research.

Chapters 2, 3, and 4 focused on identifying methods of providing feedback that are effective for student learning. Providing feedback in computer-based environments is not a new phenomenon. Nevertheless, in many learning environments, feedback interventions have already been implemented without knowledge that supports their effectiveness. Important applications of the results regarding the provision of feedback to students in CBAs include massive open online courses (MOOC). MOOCs rely heavily on computer-supported interactions between the learner and the learning environment, in which automated feedback plays a crucial role. As the use of MOOCS increases, so does the need for the systematic gathering of evidence concerning the effectiveness of methods of providing computer-based feedback. The results of this dissertation can be used to evaluate and possibly improve the feedback practices in such environments. The results also could help improve the quality of feedback provided in applications used in tablets and smart phones. Because schools are starting to use such devices for educational purposes, there is an urgent need to develop sound assessments or game-like environments in which high quality feedback is embedded.

Chapters 2, 3, and 4 investigated which methods for providing feedback are more effective in student learning. The findings showed that the effects of feedback vary for each student, and many variables influence the effect of the feedback. Future research should address how feedback can be tailored to the needs and preferences of students. Nevertheless, Smits, Boon, Sluijsmans, and Van Gog (2008) warned that students' preferences regarding feedback do not always align with what effectively increases their learning outcomes.

However, feedback does need to be perceived as relevant by students in order to be potentially effective in learning.

The results presented in Chapter 3 and 4 clearly suggested that EF is more effective than KR or KCR are in terms of increasing students' learning outcomes, particularly higher-order learning outcomes. However, the term EF has a wide range of possible meanings, and consequently the degree to which it is effective varies widely. Therefore, it would be useful to develop a classification for different methods of providing EF. Shute (2008) has provided a classification for EF that distinguishes between six types. The difficulty with this classification is that the definitions overlap, so it is sometimes hard to make an actual classification. It may be more useful to consider EF as consisting of various "building blocks" that can be built on KR or KCR. Such building blocks could cover a continuum that ranges from very subtle to very specific guidance. For example, feedback that consists of hints or strategies could provide subtle guidance, while very specific guidance could be provided in the form of detailed explanations and demonstrations of the correct solution. Although such classifications could be helpful in future studies, there is a dire need for further, high-quality research in which feedback interventions are clearly described and responsibly measured.

Furthermore, conducting experiments that investigate the effects of the stepwise provision of immediate feedback (Narciss, 2008; Stevenson, Heiser, Resing, & De Boeck, 2013), from subtle to specific, could help shed light on the effects of various methods used to provide EF. Research has suggested that the elaborateness of feedback should be adapted to the current level of the learner. Novice learners are expected to need explanations and explicit guidance, while proficient students are perhaps better provided with subtle guidance and metacognitive feedback. Similarly, novice learners could benefit the most from immediate feedback, whereas proficient learners benefit from delayed feedback. It would be very interesting to examine whether CBAs and their accompanying feedback could be optimised by taking into account item characteristics, such as the type of learning outcome, level of difficulty, and the student's ability. When adaptive adjustments of features in the learning environment can be realised, such environments could become very powerful tools for learning (Wauters, 2012). In traditional computerised adaptive tests (CAT), the selection of items is based on the estimated ability of the test taker. Items are selected to provide maximum information from a psychometrical point of view (Wainer, 2000). However, CAT algorithms aiming to measure static ability with high precision and efficiency may not be well-suited for learning environments in which feedback is provided. In these environments, the primary purpose is to support learning, not measure it. Moreover, adaptivity refers to a wide range of possible options that can be adapted to the learning or testing environment. Hence, adaptivity could also apply, for example, to adaptive item representation using multimedia, adaptive curriculum sequencing, and the adaptive provision of feedback (Wauters, 2012). Recent CAT research has made available algorithms that can be tuned to the particular purpose of the assessment (Eggen, 2013). By combining these promising developments in learning environments with CAT, further research could shed light on how to provide adaptive feedback in an effective way.

The results in the literature regarding timing are highly conflicting even though the topic has been widely studied. Chapter 4 tested the hypothesis that there is an interaction effect between feedback timing and the level of learning outcomes. The directionality of the effects was consistent with our hypothesis, showing that immediate feedback was more effective in lower-order learning outcomes and vice versa. However, statistical significance was not obtained, possibly because of the small sample size. Nevertheless, I consider this potential interaction effect worth further investigation and claim that experimental research is needed in which this specific hypothesis is tested.

Although researchers in this field are slowly becoming aware of the urgent need for studies on how educators interpret data feedback (see Chapters 5 and 6), the perspective of the student must not be forgotten. Sadler (1998), stressed that studies on how students interpret feedback and how they use the feedback for improving their learning are essential. Nevertheless, few studies have focused on the interpretation of feedback by students, and the assumption that "when students are 'given feedback' they will know what to do with it" (Sadler, 1998, p. 78) still seems to dominate. However, some studies focusing on students' feedback behaviour in CBAs have recently been conducted (e.g., Timmers &Veldkamp, 2011; Timmers, Braber-Van den Broek, & Van den Berg, 2013). The experiment reported in Chapter 2 furthermore suggested that students paid more attention to immediate feedback than to delayed feedback. I am convinced that more research is necessary in order to adapt learning environments to students' needs and demands, which would eventually lead to greater student satisfaction and higher learning outcomes.

Furthermore, the effects of feedback examined in this dissertation mainly considered the short term. Thus, there is a need for longitudinal research in order to gain insight into the effects of different methods for providing feedback over longer periods. Moreover, computer environments have a huge potential to capture data and meta-data that can be used to monitor student's progress regarding learning, as well as their (feedback) behaviour (e.g., Bouchet, Harley, Trevors, & Azevedo, 2013). One very promising line of research in this respect is data mining and associated learning analytics. This research provides rich information regarding what happens while students learn. Educators could use this information to adapt the learning environment effectively (Siemens & Long, 2011). Furthermore, students need both detailed and frequent task-related feedback and reports that provide general feedback, which could help them monitor their learning progress and stimulate active participation in their learning (Vezzu, VanWinkle, & Zapata-Rivera, 2012). In addition, the use of eye-tracking devices can provide detailed insights into student behaviour when computer-based feedback is provided. It would be worthwhile to investigate the extent to which such technologies can shed light on student behaviour in order to accommodate their needs.

In the experiments performed in this dissertation, the effects of feedback were usually measured by separate post-tests. However, this is not a very efficient method, and such methods can only discern feedback effects at a general level. In order to optimise computer-based environments for individual students, it would be worthwhile to apply or develop psychometric models that describe and explain the dynamic process of learning. I expect that such models would enable researchers to measure student learning and feedback effects more quickly and flexibly as well as to conduct experiments with such models (e.g., Stevenson et al., 2013).

Part of this dissertation focused on the interpretation of feedback by educators. This research proved worthwhile because the results suggested that users of the Computer Program LOVS encountered many difficulties in the reports and often could not interpret them correctly. However, research has suggested that the subsequent steps in the evaluative cycle, such as making decisions on how to adapt instruction, might even be more challenging (Chahine, 2013; Hattie & Brown, 2008; Heritage, Kim, Vendlinski, & Herman, 2009). Future research should point out the extent to which users are capable of transforming data feedback into meaningful educational decisions and how professional development can help overcome obstacles encountered in completing the evaluative cycle. In particular, there is an urgent need for professional development regarding the assessment literacy of teachers. Specifically, teachers seem to struggle the most in assessment-related activities, although they engage in important assessment practices every day in the classroom. Furthermore, it is unlikely that "one-shot" professional development initiatives would be sufficient. Large investments will have to be made in both the pre-service and in-service training of educational professionals if DDDM is to supporting learning. These investments should be aimed at building the necessary skills to implement formative assessment (Chapter 7) and change educators' attitudes towards formative assessment.

Furthermore, although many reports have been generated for the feedback of the results of large-scale assessments, the majority currently in use are based upon reporting for summative purposes. Consequently, the reports tend to focus on the results of learning, and not much attention is paid to the process of learning. This emphasis is not always useful in formative assessment. I believe that it would be valuable to bring a new perspective to communicating information about student learning for formative purposes. A pioneer in this area of research is Goodman. His inspiring presentation at the annual meeting of the National Council on Measurement in Education (2013) showed how this could be done by shifting away from learning outcomes and returning the focus to the student's activities and learning processes. I recommend further research that explores this formative view of reporting. Furthermore, in this same symposium session, it was discussed that the name "score report" might be inappropriate for reports that have formative purposes. The term refers to the score as an indicator for learning, which is characteristically associated with summative assessment. Furthermore, there is a need for more research on how to design reports for diagnostic tests that adequately aim to serve formative purposes. Because diagnostic tests use highly complex models, this quantitative information needs to be translated to understandable and action-oriented qualitative information.

Providing feedback to students and to educators is an essential element of formative assessment. Nevertheless, the majority of studies have focused on students. These two fields of research should be combined in one tool to achieve the maximum benefit of using computers to support formative assessment.

Although the theoretical comparison provided in Chapter 7 highlighted the promising possibilities of an overarching formative assessment approach, a crucial question remains: How to create the right balance amongst three approaches, each of which makes unique contributions to assessment practices. Furthermore, the alignment of formative with summative practices is essential in the creation of a balanced assessment system. As O'Malley, Lai, McClarty, and Way (2013) stated, "The field will need to marry information from the different assessments to produce a union that a stronger than any of the parts in isolation" (p. 146). Moreover, a balanced assessment system could lead to more efficient assessment practices because assessment results can serve multiple purposes simultaneously.

Moreover, when formative assessment is adequately implemented, the amount of time spent on remediating activities can be reduced because instruction can quickly be adapted to student needs. Several researchers have recently addressed the issue of balanced assessment systems and have proposed frameworks that could help shape such systems (Brookhart, 2013; Stiggins & Chappuis, 2013). Nevertheless, the explicit functions of the three approaches to formative assessment within such an assessment system, and most importantly, the practical implications of such frameworks for a balanced assessment systems, should be addressed in future research.

References

Bouchet, F., Harley, J. M., Trevors, G. J., & Azevedo, R. (2013). Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning. *Journal of Educational Data Mining, 5*(1), 104–146. Retrieved from http://www.educationaldatamining.org/JEDM/ index.php?option=com_content&view=article&id=96&Itemid=81

Brookhart, S. M. (2013). Comprehensive assessment systems in service of learning. In R. W. Lissitz (Ed.), *Informing the practice of teaching using formative and interim assessment: A systems approach* (pp.165–184). Charlotte, NC: Information Age.

Chahine, S. (2013, April). *Investigating educators' statistical literacy and score report interpretation.* Paper presented at the conference of the National Council on Measurement in Education, San Francisco, CA.

Goodman, D. (2013, April). *Communicating student learning—one jurisdiction's efforts to change how student learning is reported.* Paper presented at the conference of the National Council on Measurement in Education, San Francisco, CA.

Eggen, T. J. H. M. (2013, August). *Multi segment computerized adaptive testing*. Paper presented at the EARLI conference, München, Germany.

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. doi:10.2190/ET.36.2.g

Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement: Issues and Practice*, *28*, 24–31. doi:10.1111/j.1745-3992.2009.00151.x

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merril, J. J. G. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 125–144). Mahwah, NJ: Lawrence Erlbaum Associates.

O'Malley, K., Lai, E., McClarty, K., & Way, D. (2013). Marrying formative, periodic, and summative assessments: I do. In R. W. Lissitz (Ed.), *Informing the practice of teaching using formative and interim assessment: A systems approach* (pp. 145–164). Charlotte, NC: Information Age.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119–144. doi:10.1007/BF00117714

Sadler, D. R. (1998). Formative Assessment: Revisiting the territory. *Assessment in Education: Principles, Policy & Practice*, *5*, 77–84. doi:10.1080/0969595980050104

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153–189. doi:10.3102/0034654307313795

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review, 46*(5), 30–32. Retrieved from http://www.elmhurst.edu/~richs/EC/OnlineMaterials/SPS102/Teaching%20and%20Learning/Penetrating%20the%20Fog.pdf

Smits, M., Boon, J., Sluijsmans, D. M. A., & Van Gog, T. (2008). Content and timing of feedback in a web-based learning environment: Effects on learning as a function of prior knowledge. *Interactive Learning Environments, 16*, 183–193. doi:10.1080/10494820701365952

Stevenson, C. E., Heiser, W. J. H., Resing, W. C. M., & De Boeck, P. A. L. (2013, July). *Individual differences in the effect of feedback on children's change in analogical reasoning.* Paper presented at the international workshop formative feedback in interactive learning environments, Memphis, TN. Retrieved from http://ceur-ws.org/Vol-1009/0806.pdf

Stiggins, R., & Chappuis, S. (2013). Productive formative assessment always requires local district preparation. In R. W. Lissitz (Ed.), *Informing the practice of teaching using formative and interim assessment: A systems approach* (pp.237–247). Charlotte, NC: Information Age.

Timmers, C. F., Braber-Van den Broek, J., & Van den Berg, S. M. (2013). Motivational beliefs, student effort, and feedback behaviour in computer-based formative assessment. *Computers & Education*, *60*, 25–31. doi:10.1016/j.compedu.2012.07.007

Timmers, C. F., & Veldkamp, B. P. (2011). Attention paid to feedback provided by a computer-based assessment for learning on information literacy. *Computers & Education, 56,* 923–930. doi:10.1016/j.compedu.2010.11.007

Vezzu, M., VanWinkle, W., & Zapata-Rivera, D. (2012). *Designing and evaluating an interactive score report for students*. Princeton, NJ: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/RM-12-01.pdf

Wainer, H. (Ed.) (2000). *Computerized adaptive testing. A primer* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Wauters, K. (2012). *Adaptive item sequencing in item-based learning environments* (Unpublished doctoral dissertation). Katholieke Universiteit Leuven, Leuven, Belgium).

Summary

Formative assessment concerns any assessment that provides feedback that is intended to support learning and can be used by teachers and/or students. Computers could offer a solution to overcoming obstacles encountered in implementing formative assessment. For example, computer-based assessments could be scored automatically, students could be provided with automatically generated feedback, and reports that provide information on student learning could be generated automatically for use by teachers and others. The extent to which computers could support formative assessment practices was the central focus of this dissertation, which covered three main areas:

1. Item-based feedback provided to students through a computer (Chapters 2, 3, and 4)
2. Feedback provided through a computer to educators based on students' assessment results (Chapters 5 and 6)
3) A comparison of three approaches to formative assessment: Data-based decision making (DBDM), assessment for learning (AfL), and diagnostic testing (DT) (Chapter 7)

Chapter 1 provided a general introduction. A conceptual framework elaborated the broader context of this dissertation—formative assessment. The following working definition of formative assessment was used in this dissertation: *Any assessment that provides feedback that is intended to support learning and can be used by teachers and/or students.* Furthermore, the characteristics of feedback provided to both students and educators were discussed. The last part of the conceptual framework focused on the use of computers in educational assessment.

In Chapter 2, the effects of written feedback in a computer-based assessment for learning on students' learning outcomes were investigated in an experiment at an institute of higher education in the Netherlands. Students were randomly assigned to three groups and then were subjected to an assessment for learning that used different kinds of feedback: immediate knowledge of correct response (KCR) + elaborated feedback (EF); delayed KCR + EF; and delayed knowledge of results (KR). A summative assessment was used as a post-test. No significant effect was found in the post-test of the feedback condition on student achievement. The results suggested that students paid more attention to immediate than to delayed feedback. Furthermore, the time spent reading feedback was positively influenced by students' attitude and motivation. Students perceived immediate KCR + EF to be more useful in learning than KR. Students also had a more positive attitude towards feedback in a CBA when they received KCR + EF instead of KR only.

Chapter 3 presented a review of the relevant literature regarding the effectiveness of different methods in providing written feedback through a computer-based assessment for learning. In analysing the results, a distinction was made between lower-order learning outcomes (e.g., recalling and understanding) and higher-order learning outcomes (e.g., the application of knowledge and skills in new situations). The few available high-quality studies suggested that students could benefit from KCR to obtain lower-order learning outcomes. In addition, EF seems beneficial for gaining both lower-order and higher-order learning

outcomes. Furthermore, the literature reviewed showed that a number of variables should be taken into account when investigating the effects of feedback on learning outcomes.

Chapter 4 presented a meta-analysis of the effects of methods used to provide item-based feedback on students' learning outcomes in a computer-based environment. Seventy effect sizes were computed from 40 studies, which ranged from -0.78 to 2.29. The results showed that EF, e.g., feedback that provides an explanation, produced higher effect sizes (0.49) than feedback regarding the correctness of the answer (KR; 0.05) or providing the correct answer (KCR; 0.32). EF was particularly more effective than KR and KCR for higher-order learning outcomes. The effect sizes were positively affected by EF. Larger effect sizes were found for mathematics compared to social sciences, science, and languages. The effect sizes were negatively affected by delayed feedback timing, and primary and high school settings (as opposed to higher education). Although the results suggested that immediate feedback was more effective for lower-order learning than delayed feedback was, and vice versa, no significant interaction was found.

Chapter 5 focused on the interpretation of score reports generated by the Computer Program LOVS by teachers, internal support teachers, and principals. This study was conducted from a data-driven decision making (DDDM) perspective (called data-based decision making [DBDM] in the recent literature). DDDM, such as the decision making that is conducted by using pupil-monitoring systems, has become increasingly popular in the Netherlands because it is considered to have potential 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. The study performed in this chapter aimed to investigate the extent to which the reports are correctly interpreted by educators and to identify various potential obstacles to the interpretation of the reports. The results suggested that users encounter many difficulties in these reports and often cannot interpret them entirely correctly. An important lesson is that although the reports generated by the Computer Program LOVS have been in use for a couple of years, many users struggle with interpreting them. We concluded that it seemed worthwhile to examine whether redesigned score reports would be interpreted more correctly.

Chapter 6 investigated how the reports generated by the Computer Program LOVS could be redesigned to support users in interpreting pupils' test results. In several rounds of consultations with users and experts, alternative designs for the reports were created and field tested. No clear differences were found in the accuracy of users' interpretations between the original and the redesigned versions of the reports. However, users' perceptions of the redesigned reports were predominantly positive. Eventually, a final set of design solutions was generated. The authors emphasise the need for clear score reports and the involvement of experts as well as current and future users in the design process in order to ensure the validity of the reports.

Chapter 7 was not directly concerned with computer-based feedback in formative assessment but related to a broader approach in which computer-based applications may be used to support learning. The chapter provided a theoretical comparison of three approaches to formative assessment, which are currently the most frequently discussed in the literature on educational research: data-based decision making (DBDM); assessment for learning (AfL); and diagnostic testing (DT). The differences and similarities in the implementation of each approach were explored. The results revealed that although differences exist amongst the theoretical underpinnings of DBDM, AfL, and DT, the implementation of an overarching formative assessment and formative evaluation approach is possible. Moreover, the integration of the three assessment approaches can lead to formative decisions that are more valid compared to those made at present. Future research is needed to examine the actual implementation of a combination of these three approaches, including their associated challenges and opportunities.

Samenvatting

Formatief assessment heeft betrekking op assessment dat feedback geeft welke bedoeld is om leren te ondersteunen en gebruikt kan worden door leerkrachten en/of studenten. Computers kunnen mogelijk een oplossing bieden bij het overwinnen van obstakels bij de implementatie van formatief assessment. Enkele voorbeelden hiervan zijn computergestuurde assessments die automatisch gescoord kunnen worden, computers kunnen studenten automatisch gegenereerde feedback geven en rapportages welke informatie geven over het leren van studenten kunnen automatisch worden gegenereerd voor het gebruik door, onder andere, leerkrachten.

De mate waarin de computer ondersteuning kan bieden bij formatief assessment was de centrale focus van deze dissertatie, welke grofweg drie thema's beslaat:

1. Item-gerelateerde computergestuurde feedback gegeven aan studenten (Hoofdstuk 2, 3 en 4);
2. Feedback over toetsresultaten van studenten aan onderwijspersoneel (zoals leerkrachten en directeuren) via een computer (Hoofdstuk 5 en 6); en
3. Een vergelijking van drie benaderingen naar formatief assessment: data-based decision making (DBDM; in de Nederlandstalige literatuur aangeduid als opbrengstgericht werken [OWG]), assessment for learning (AfL; in Nederland ook wel aangeduid als assessment voor het leren) en diagnostisch toetsen (DT) (Hoofdstuk 7).

In Hoofdstuk 1 is een algemene introductie gegeven. Hierin is een conceptueel raamwerk gepresenteerd waarin werd ingegaan op de bredere context van deze dissertatie—formatief assessment. De volgende werkdefinitie van formatief assessment is gebruikt in deze dissertatie: *elk assessment dat feedback geeft welke bedoeld is om leren te ondersteunen en gebruikt kan worden door leerkrachten en/of studenten.* Verder worden de kenmerken van feedback naar studenten en de kenmerken van feedback naar onderwijspersoneel besproken. Het laatste deel van het conceptuele raamwerk was gericht op het gebruik van computers in assessments in onderwijssituaties.

In Hoofdstuk 2 zijn de effecten van geschreven feedback in een computergestuurde formatieve toets onderzocht. Voor dit onderzoek is een experiment uitgevoerd op een HBO-instelling in Nederland. Studenten zijn aselect toegewezen aan drie groepen, welke ieder een andere vorm van feedback ontvingen. Deze vormen van feedback zijn: 1) directe kennis van de correcte respons (knowledge of correct response; KCR) + uitgebreide feedback (elaborated feedback; EF), 2) uitgestelde KCR + EF en 3) uitgestelde kennis van het resultaat (knowledge of results; KR). Een summatieve toets is gebruikt als post-test. Er zijn geen significante effecten gevonden van de feedback condities op de leeropbrengsten van studenten op de post-test. Resultaten suggereren dat studenten meer aandacht besteedden aan directe dan aan uitgestelde feedback. Verder bleek de tijd die studenten besteedden aan het lezen van feedback positief gerelateerd te zijn aan hun attitude en motivatie. Studenten vonden feedback in de vorm van KCR + EF nuttiger ter ondersteuning van het leren dan KR. Bovendien hadden studenten die KCR + EF ontvingen een positievere attitude ten opzichte van de computergestuurde formatieve toets dan de studenten die alleen KR ontvingen.

In Hoofdstuk 3 is een systematische reviewstudie gepresenteerd. Deze review had tot doel op basis van de beschikbare literatuur een overzicht te geven van de effecten van verschillende methoden om geschreven feedback te geven in computergestuurde formatieve assessments. In de analyse van de resultaten is onderscheid gemaakt tussen lagere-orde leeropbrengsten (zoals onthouden en begrijpen) en hogere-orde leeropbrengsten (zoals het toepassen van kennis en vaardigheden in nieuwe situaties). Het kleine beetje onderzoek van hoge kwaliteit dat beschikbaar is suggereert dat studenten kunnen profiteren van KCR voor het verwerven van lagere-orde leeropbrengsten. Verder blijkt EF bevorderlijk voor het verwerven van zowel lagere-orde als hogere-orde leeropbrengsten. De resultaten van deze review suggereren dat verschillende variabelen in acht moeten worden genomen bij het onderzoeken van de effecten van feedback op leren.

Hoofdstuk 4 betrof een meta-analyse naar de effecten van verschillende methoden voor het geven van item-gerelateerde feedback in een computergestuurde omgeving op de leeropbrengsten van studenten. Op basis van 40 studies zijn 70 effecten berekend, welke varieerden van -0,78 tot 2,29. De resultaten wijzen uit dat EF, zoals het geven van uitleg, leidt tot hogere effecten (0,49) dan feedback welke betrekking heeft op de correctheid van het antwoord (KR; 0,05) of het geven van het juiste antwoord (KCR; 0,32). EF was in het bijzonder meer effectief dan KR en KCR voor hogere-orde leeropbrengsten. Effecten waren positief beïnvloed door het feedbacktype EF. Verder werden de hogere effecten gevonden voor wiskunde in vergelijking met sociale wetenschappen, natuurwetenschappen en talen. Effecten werden negatief beïnvloed door uitgestelde feedback en basisonderwijs + middelbaar onderwijs (ten opzichte van hoger onderwijs). Alhoewel de resultaten suggereerden dat directe feedback meer effectief is voor lagere-orde leeropbrengsten dan uitgestelde feedback en vice versa, is er geen significant interactie-effect gevonden.

Hoofdstuk 5 was gericht op de interpretatie van rapportages van het Computerprogramma LOVS door leerkrachten, intern begeleiders en directeuren. Dit onderzoek was uitgevoerd vanuit het perspectief van data-driven decision making (DDDM, ook wel data-based decision making [DBDM] genoemd), in Nederland aangeduid als opbrengstgericht werken (OGW). Recentelijk heeft opbrengstgericht werken aan populariteit gewonnen in Nederland. Het wordt namelijk gezien als een veelbelovende manier om de leeropbrengsten van leerlingen te verhogen. Leerlingvolgsystemen kunnen hierbij zeer waardevol zijn. De rapportages uit het Computerprogramma LOVS (Cito), behorende bij het leerlingvolgsysteem LOVS, verschaffen leerkrachten betrouwbare en objectieve datafeedback. Echter, onderzoek heeft aangetoond dat veel gebruikers moeite hebben met de interpretatie van deze rapportages. Het onderzoek in dit hoofdstuk was bedoeld om na te gaan in hoeverre de rapportages correct geïnterpreteerd worden door de verschillende groepen gebruikers en om verschillende mogelijke struikelblokken te identificeren met betrekking tot de interpretatie van de rapportages. De resultaten wijzen uit dat er verschillende struikelblokken zijn voor gebruikers bij de interpretatie van de rapportages van het Computerprogramma LOVS en dat gebruikers vaak niet in staat zijn deze rapportages volledig correct te interpreteren. Een belangrijke les uit dit onderzoek is dat ondanks dat de rapportages van het Computerprogramma LOVS al jaren in gebruik zijn, veel gebruikers moeite hebben met de interpretatie van deze rapportages. We concludeerden dat het de moeite waard leek om te onderzoeken of herontworpen rapportages kunnen leiden tot meer correcte interpretaties.

In Hoofdstuk 6 is onderzocht hoe de rapportages van het Computerprogramma LOVS kunnen worden herontworpen, op een manier die gebruikers ondersteunt bij het interpreteren van de toetsresultaten van leerlingen. In verschillende ronden van consultatie en ontwerp in samenwerking met gebruikers en experts, zijn alternatieve ontwerpen voor de rapportages gecreëerd en uitgetest. Er zijn geen duidelijke verschillen gevonden met betrekking tot de accuraatheid van interpretaties door gebruikers tussen de originele en herontworpen versies van de rapportages. De percepties van gebruikers over de herontworpen rapportages waren echter voornamelijk positief. We benadrukten de behoefte aan duidelijke rapportages en de noodzaak om experts en huidige/toekomstige gebruikers te betrekken bij het ontwerpproces, om de validiteit van de rapportages te waarborgen.

Hoofdstuk 7 was niet direct gerelateerd aan computergestuurde feedback in formatieve assessments, maar was gericht op de bredere benadering waarin computergestuurde toepassingen gebruikt kunnen worden om leren te ondersteunen. Dit hoofdstuk presenteerde een theoretische vergelijking van drie benaderingen naar formatief assessment, die tegenwoordig het meest bediscussieerd worden in de onderwijskundige literatuur: data-based decision making (DBDM; in de Nederlandstalige literatuur opbrengstgericht werken [OWG] genaamd), assessment for learning (AfL; in Nederland ook wel aangeduid als assessment voor het leren) en diagnostisch toetsen (DT). Verder zijn de verschillen en overeenkomsten in de implementatie van elke benadering verkend. Dit onderzoek wijst uit dat ondanks verschillen in de theoretische basis van DBDM, AfL en DT, er mogelijkheden zijn voor het implementeren van een overkoepelende benadering naar formatief assessment en formatieve evaluatie. De integratie van deze drie benaderingen naar assessment, kan bovendien leiden tot meer valide formatieve beslissingen. Vervolgonderzoek is nodig om te onderzoeken hoe een daadwerkelijke implementatie van een mix van deze benaderingen, elk met zijn eigen uitdagingen en mogelijkheden, kan worden gerealiseerd.

Research Valorisation: Publications and Presentations

1. Publications in Scientific Journals

1.1 Published

Van der Kleij, F. M. & Eggen, T. J. H. M. (2013). Interpretation of the score reports from the Computer Program LOVS by teachers, internal support teachers and principals. *Studies in Educational Evaluation, 39*, 144–152. doi:10.1016/j.stueduc.2013.04.002

Van der Kleij, F. M., Eggen, T. J. H. M., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers & Education, 58*, 263–272. doi:10.1016/j.compedu.2011.07.020

Van der Kleij, F. M., Timmers, C. F., & Eggen, T. J. H. M. (2011). The effectiveness of methods for providing written feedback through a computer-based assessment for learning: A systematic review. *CADMO, 19*, 21–39. doi:10.3280/CAD2011-001004

1.2 Submitted

Van der Kleij, F. M., Eggen, T. J. H. M., & Engelen, R. J. H. (submitted). *Towards valid score reports in the Computer Program LOVS: A redesign study.* Manuscript submitted for publication.

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (submitted). *Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis.* Manuscript submitted for publication.

Van der Kleij, F. M., Vermeulen, J. A., Schildkamp., K., Eggen, T. J. H. M. (submitted). *Data-based decision making, assessment for learning, and diagnostic testing in formative assessment*. Manuscript submitted for publication.

2. Professional Publications

Timmers, C. F., & Van der Kleij, F. M. (2012). De formative functie van toetsen. *Examens, 9*(3), 5–9.

Van der Kleij, F. M., & Timmers, C. F. (2011). *Leeropbrengst van feedback in computergestuurde toetsen; wat is effectief?* Available from http://195.169.48.3/html/feedback/Feedback_in_CBA.pdf

Van der Kleij, F. M., Vermeulen, J., Eggen, T. J. H. M., & Veldkamp, B. P. (2012). *Leren van toetsen; een cyclisch proces.* Available from http://toetswijzer.kennisnet.nl/html/leren_van_toetsen/leren_van_toetsen.pdf

3. Book Chapters

Vermeulen, J. A., & Van der Kleij, F. M. (2012). Towards an integrative formative approach of data-driven decision making, assessment for learning, and diagnostic testing. In T. J. H. M. Eggen, & B. P. Veldkamp (Eds.), *Psychometrics in practice at RCEC* (pp. 151–179). Enschede, the Netherlands: Ipskamp. doi:10.3990/3.9789036533744

4. Conference Contributions, Presentations, and Lecturers

Timmers, C. F., & Van der Kleij, F. M. (2011, November). *Feedback in formative CBAs to support learning.* Poster presented at the conference of the European Association for Educational Assessment, Belfast, Northern Ireland.

Van der Kleij, F. M. (2011, June). *Computergestuurd Assessment for learning; inzichten in de effectiviteit van feedback op leren.* Presentation at Cito's yearly math day. Arnhem, the Netherlands.

Van der Kleij, F. M. (2012, November).*Van toetsresultaten naar inzicht in leeropbrengsten: interpretatie van de rapportages uit het Computerprogramma LOVS.* Paper presented at the Onderwijssummit, Utrecht, the Netherlands.

Van der Kleij, F. M. & Eggen, T. J. H. M. (2012, June). *Van toetsresultaten naar inzicht in leeropbrengsten: Interpretatie van de LOVS rapportages door leerkrachten.* Paper presented at the Onderwijs Research Dagen, Wageningen, the Netherlands.

Van der Kleij, F. M., Eggen, T. J. H. M., & Engelen, R. J. H. (2012, October). *Towards valid score reports in the Computer Program LOVS: A redesign study.* Paper presented at the conference of the Research Centre for Examinations and Certification, Enschede, the Netherlands.

Van der Kleij, F. M., Eggen, T. J. H. M., & Engelen, R. J. H. (2013, May). *Naar valide rapportages in het Computerprogramma LOVS: Een herontwerp studie.* Paper presented at the Onderwijs Research Dagen, Brussels, Belgium.

Van der Kleij, F. M., Eggen, T. J. H. M., Timmers, C. F., & Veldkamp, B. P. (2010, November). *Effectiveness of feedback in a computer-based assessment for learning.* Paper presented at the conference of the European Association for Educational Assessment, Oslo, Norway.

Van der Kleij, F. M., Eggen, T. J. H. M., Timmers, C. F., & Veldkamp, B. P. (2011, June) *Effectiveness of feedback in a computer-based assessment for learning.* Paper presented in a webinar for members of the European Association for Educational Assessment.

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2012, November). *Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis.* Paper presented at the conference of the European Association for Educational Assessment, Berlin, Germany.

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2013, April). *Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis.* Paper presented at the conference of the National Council on Measurement in Education, San Francisco, CA.

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2013, June). *Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis.* Paper presented at the conference of the International Association for Cognitive Education and Psychology, Leiden, the Netherlands.

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2013, October). *Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis.* Paper presented at the conference of the International Association for Educational Assessment, Tel Aviv, Israel.

Van der Kleij, F. M., Timmers, C. F., & Eggen, T. J. H. M. (2011, June). *De effectiviteit van verschillende manieren om geschreven feedback te geven in een computergestuurd assessment op leeropbrengsten van studenten: Een systematische review.* Paper presented at the Onderwijs Research Dagen, Maastricht, the Netherlands.

Van der Kleij, F. M., Timmers, C. F., & Eggen, T. J. H. M. (2011, November). *The effectiveness of methods for providing written feedback through a computer-based assessment for learning: A systematic review.* Paper presented at the conference of the European Association for Educational Assessment, Belfast, Northern Ireland.

Van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., Eggen, T. J. H. M. (2013, November). *Data-based decision making, assessment for learning, and diagnostic testing in formative assessment.* Keynote presentation at the conference of the European Association for Educational Assessment, Paris, France.

Visscher, A. J., & Van der Kleij, F. M. (2013, March). *The Cito student monitoring system; features and what it takes to benefit from it* (Lecture 5, Assessment of and for Learning, Master Educational Science and Technology). University of Twente, Enschede, the Netherlands.

Acknowledgements

I would like to thank everyone who supported and inspired me during this Ph.D. project, whether consciously or unconsciously. I appreciate the importance of timely and elaborated feedback in a learning process. In particular, I thank my promoter and supervisor, Theo Eggen. His great support during this research project was invaluable. His adaptive supervision style suited me, and I am thankful for our collaboration and shared moments of fun. Although he is also new to the field of (computer-based) formative assessment, he was always enthusiastic about my research, eager to learn, and to make me learn as much as possible. I would also like to thank Theo for reading many texts, sometimes on very short notice. The most serious disagreement we had about a manuscript was solved by placing a comma in a sentence. Although in the early stages, the results of my project were not clearly defined, his feedback always gave me direction. Thank you for allowing me to choose my own path and supporting me to the end.

I thank Caroline Timmers for the wonderful discussions, beneficial collaboration, and the facilitation of the experiment at Saxion University of Applied Sciences. I would also like to thank Bernard Veldkamp for suggesting the research topic of feedback in computer-based formative assessments as well as his encouraging enthusiasm and continual interest in my research.

I am grateful to Cito and RCEC for facilitating this Ph.D. project and letting me explore this new research direction. I am also grateful for the countless opportunities I have been given in my academic development, particularly the collaborations with others, both within and outside Cito. I also express my gratitude to my colleagues at Cito and the University of Twente, particularly those at RCEC, for helping to create a positive working climate and informal atmosphere. I would especially like to thank Jorine Vermeulen for the many inspirational talks over coffee, and of course for the wonderful week of academic writing in the UK. In addition, I also thank Anton Béguin for giving me the opportunity to conduct this research at POK and for initiating the idea of conducting a meta-analysis. Despite the copious amounts blood, sweat, and tears that I shed, it was worthwhile. I also thank Birgit for her everlasting cheerfulness and how she took care of many things, particularly the formatting of my dissertation. Furthermore, I would like to thank Remco Feskens, Ronald Engelen, and Kim Schildkamp for their wonderful contributions to papers incorporated in this dissertation. I also thank the members of the graduation committee for their time and valuable judgement.

I also thank Jaqueline Visser and Geert Evers for letting me do research on the reports of the Computer Program LOVS and then redesign the reports later on. Furthermore, I extend my gratitude to Gerben Veerbeek for assisting in the focus group meetings. I also thank the experts Ilse Papenburg, Ilonka Verheij, Judith Vos, and Laura Staman. I also thank Servaas Frissen and Meike Köhler for their very much appreciated research assistance. I of course thank all those who participated in the experiments and field consultations that were performed to gather the research data for this dissertation.

Finally, I thank my family and friends for their support and the much-needed moments of fun and distraction. I would like to thank my friends, Eline Roelofs and Merijn Halfers, for assisting and supporting me as paranymphs during the defence of this dissertation. I thank Jamila for designing the cover of my dissertation. I give special thanks to my husband Chris, who gave me the time to take a detour before realising our dream of moving to Australia together. Thank you for your patience, loving support, encouragement, and constant presence, no matter the distance.