Computers & Education 58 (2012) 263–272

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

# Effects of feedback in a computer-based assessment for learning

Fabienne M. van der Kleij [a,*], Theo J.H.M. Eggen [a,b,1], Caroline F. Timmers [c], Bernard P. Veldkamp [b,2]

[a] Cito, PO Box 1034, 6801 MG Arnhem, The Netherlands
[b] University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands
[c] Saxion University of Applied Sciences, PO Box 70.000, 7500 KB Enschede, The Netherlands

ARTICLE INFO

*Article history:*
Received 26 April 2011
Received in revised form
22 July 2011
Accepted 30 July 2011

*Keywords:*
Evaluation of CAL systems
Media in education
Post-secondary education
Teaching/learning strategies
Evaluation methodologies

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 influenced by students' attitude and motivation. Students perceived immediate KCR + EF feedback 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. 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 didactic 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, 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. 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

* Corresponding author. Tel.: +31263521599.
*E-mail addresses:* fabienne.vanderkleij@cito.nl (F.M. van der Kleij), theo.eggen@cito.nl (T.J.H.M. Eggen), c.f.timmers@saxion.nl (C.F. Timmers), b.p.veldkamp@utwente.nl (B.P. Veldkamp).
[1] Tel.: +31263521468.
[2] Tel.: +31534893653.

0360-1315/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compedu.2011.07.020

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). 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, & van Gog, 2008; Wang, 2011). In other studies, no effects were found (Clariana & Lee, 2001; 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.

## 3. 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.

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) (Fig. 1). The type and level of feedback together determine the content of the feedback. KR and KCR feedback 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 Fig. 1. Besides the content of the feedback, timing (Shute) plays an important role. Feedback can be provided either immediately or with delay.

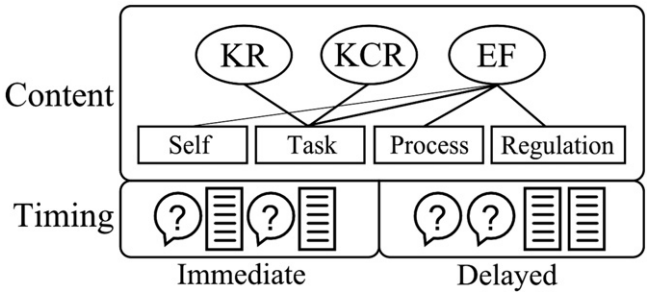


**Fig. 1.** Types of feedback distinguished by Shute (2008) linked to levels of feedback distinguished by Hattie and Timperley (2007) and timing (Shute).

## 4. 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 Fig. 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). 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). 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), and subsequently make use of it.

## 5. 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 feedback 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 feedback (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 A). 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.*

## 6. Method

### 6.1. Participants

A group of 152 first-year students ($N = 152$, 28 women, 124 men, $M_{age} = 20.30$ years; $SD = 2.45$, age range: 17–27 years) in Commercial Economics (CE) at an institute for higher education participated in this study. The students were selected from nine classes. These students

**Table 1**
Feedback conditions within the experiment.

| Group 1 | Group 2 | Group 3 |
|---|---|---|
| Immediate KCR + elaborated feedback | Delayed KCR + elaborated feedback | Delayed KR |
| Immediate computer-based written feedback: Knowledge of correct response. | Delayed (after completing the entire assessment for learning) computer-based written feedback: Knowledge of correct response. | Delayed (after completing the entire assessment for learning) computer-based written feedback: Knowledge of results. |
| 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. |

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.

### 6.2. 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.

### 6.3. 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 A, and a screenshot of the summative assessment can be found in Appendix B. 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.[3] 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.

#### 6.3.1. 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 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.

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

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 1.

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 C.

#### 6.3.2. 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.

#### 6.3.3. 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.

#### 6.3.4. 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.

#### 6.3.5. 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.

#### 6.3.6. 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 $\alpha = .85$. The summative assessment was judged to be of insufficient quality. For this assessment, $\alpha = .40$. 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 $\alpha = .66$, 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 ($\theta$) 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 result of the CTT and IRT analyses also showed that the summative assessment was more difficult than the assessment for learning.

### 6.4. 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 min. These restrictions were put in place to make sure all students would be seriously engaged in the CBA.

### 6.5. 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 ($\alpha$) 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.

## 7. Results

### 7.1. 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 shows the results of the comparisons of the mean proportions correct for the three groups.

From Table 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, $F(2, 149) = 13.99$, $p = .822$, $\eta^2 = .003$, 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.

**Table 2**
Proportion correct on the assessment for learning and summative assessment and time in seconds spent reading feedback.

| Group | Proportion correct Assessment for learning | | Proportion correct Summative assessment | | Time spent reading feedback | |
|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Group 1[a] | .62 | .20 | .63 | .22 | 257.60 | 129.38 |
| Group 2[b] | .65 | .19 | .66 | .21 | 138.60 | 149.09 |
| Group 3[c] | .68 | .21 | .65 | .22 | 96.61 | 75.30 |
| Total[d] | .65 | .20 | .65 | .22 | 165.57 | 138.95 |

[a] $n = 52$.
[b] $n = 47$.
[c] $n = 51$.
[d] $N = 150$.

**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 learnt 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.

### 7.2. 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.

Table 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.

### 7.3. 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.

### 7.4. 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 influenced the time spent reading feedback, which implies Hypothesis 3 was not rejected.

## 8. 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; 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.

## Appendix. Supplementary data

Supplementary data related to this article can be found online at doi:10.1016/j.compedu.2011.07.020.

## References


Assessment Reform Group. (1999). *Assessment for learning: Beyond the black box*. Cambridge University.

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research, 13*(2), 111–127.

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*(2), 213–238.

Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dory, Y., Ridgway, J., et al. (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*(3), 23–36.

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.

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 & Education, 39*(2), 183–200.

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: 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(2), 254–284.

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*, 1–20. doi:10.1007/s11423-010-9153-6.

Lopez, L. (2009). Effects of delayed and immediate feedback in the computer-based testing environment [Electronic version]. Doctoral dissertation, Department of Curriculum, Instructional, and Media Technology, Indiana State University. Available from http://gradworks.umi.com/33/58/3358462.html.

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.

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.

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

Miller, T. (2009). Formative computer-based assessments: the potentials and pitfalls of two formative computer-based assessments used in professional learning programs. *Dissertation Abstracts International, 70*, 4,, AATNR48227. (9780494482278).
Question Mark Perception (Version 4.0) [Computer software]. Available from http://www.questionmark.co.uk/us/index.aspx.
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: 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*. London: 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: 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.
Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). *OPLM: One parameter logistic model (Version 3.0) [Computer software]*. Arnhem: 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, Inc.
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.