

Instituut ELAN
Instituut voor
Expertise-ontwikkeling in het VO
Lerarenopleiding
Aansluiting VO-HO
Nascholing in het VO



DRAFT

The effect of pre-test sensitizing in a digital system on the acquisition of science concepts

Floris A.B.H. Bos & Cees Terlouw, University of Twente, Enschede, the Netherlands
Albert Pilot, University of Utrecht, the Netherlands
(2007)

Paper presented at the 12th Biennial Conference of the
European Association for Research on Learning and Instruction,
EARLI 2007,
Budapest 27 August – 2 September 2007

Contact:
Floris A.B.H. Bos
University of Twente
abh.bos@home.nl

GW-ELAN.07-3356

Universiteit Twente
GW - ELAN
Postbus 217
7500 AE ENSCHEDE

tel.: 053 - 489 3560
fax: 053 - 489 4755
elan@edte.utwente.nl
<http://www.utwente.nl/elan>

The effect of pre-test sensitizing in a digital system on the acquisition of science concepts

Floris A.B.H. Bos & Cees Terlouw, University of Twente (Enschede, the Netherlands), Albert Pilot, University of Utrecht (the Netherlands) (2007)

Index:

ABSTRACT	5
1 RATIONALE	5
2 THEORETICAL FRAMEWORK	5
3 RESEARCH QUESTIONS	8
4 METHOD	8
5 RESULTS	15
6 CONCLUSIONS	16
7 DISCUSSION	16
8 APPENDIX 1	18
9 REFERENCES	19

Abstract

In an experimental instructional design the effect of sensitizing with a pre-test consisting of two different question types was tested. The pre-test was embedded in an interactive digital system giving an orientation on science concepts in the joint area of physics, chemistry, biology, applied mathematics, and computer sciences. An extended Solomon Four Group design was used: science concepts were made operational or not, a sensitizing pre-test was given or not, and the pre-test consisted of short-answer questions or multiple-choice questions. The results showed high learning gains, especially after applying a pre-test in combination with making the science concepts operational. The pre-test effect did not depend on the question type. Applying a pre-test without making the science concepts operational had no significant learning effect.

1 Rationale

A new pre-university curriculum has been implemented with a specific view on the learning process in the Dutch secondary education at the end of the millennium. The teacher role in this new curriculum has more or less been changed from source of knowledge and certifying agent to that of a coach in a process of learning by personal discovery and exploration in a self-directed and often in a cooperative setting. However, because of this role change and, at the same time, a decrease in study load, also the availability of a subject teacher in chemistry, mathematics and physics for students has been decreased dramatically (Tweede Fase Adviespunt, 2005).

Since this decrease in teacher availability is structural (Ritzen, 2006; Roes, 2001) it is desirable to do research on the increase of the effectiveness of efficient learning methods. The application of ICT in optimal settings is obvious and promising for beta-sciences in this respect (Osborne & Hennessy, 2003).

In our former research on peer support in combination with discovery learning in a computer simulation based environment, a pre-test was used as an indication for entering behaviour (Bos, Terlouw, & Pilot, 2007a). In this particular experiment the electronic learning environment showed to give a significantly high learning gain, but noteworthy enough the effect of the pre-test was even higher. This result suggested that the pre-test effect could be used in a systematic way to increase effectively the output of an ICT based educational arrangement. According to our explanation the sensitizing effect of a pre-test concerns the activation of prior knowledge that facilitates the next learning of new knowledge. Applying the insight that from an educational perspective pre-testing is not a problem but an opportunity lead to the question of what kind of question should be asked in a pre-test. In the research project mentioned above open questions were used. In a computerized environment the use of closed questions however are implemented much more easily, especially when immediate feedback has to be given. Pre-testing with closed questions in an ICT-environment could be a prospective and easy to be implemented extension to the repertory of methods that activate prior knowledge such as concept mapping, brainstorming, etc. The use of such an ICT-environment should be investigated in an ecological (classroom) situation and, if possible, for more than one subject. In the current research it was decided to investigate the effect in a digital introduction, including a sensitizing pre-test, to lessons in General Science, Chemistry and Computer Science at the beginning of the upper level secondary school, class 4 of a six year pre-university curriculum (in Dutch terms: 4VWO).

2 Theoretical framework

This article is about learning new science concepts and its concomitant accretion of the already existing conceptual network. It is important that prior to the actual acquisition, relevant schemes in long-term memory are activated in order to connect with new knowledge.

In terms of ACT* theory this could be called a transfer of existing schemes from long term memory to the more easily accessible short term memory. In this way the adaptation of the existing network is more easy and effective (Anderson & Schunn, 2000).

Actually, Anderson & Schunn (2000) describe the process of the activation of prior knowledge. It is difficult to overestimate the role of prior knowledge in learning. Dochy, Segers, & Buehl, (1999)

surveyed thoroughly the role of prior knowledge and the influence of the assessment method of prior knowledge:

- (a) There is a strong relation between prior knowledge and performance: 92% of the 183 reviewed studies report positive effects. Between 30 and 60% of the variance is explained by prior knowledge. Next to this, other variables are important such as learning strategy, meta-cognitive knowledge, interest in the subject matter and beliefs. Other factors are availability, access to and structure of the knowledge. However, misconceptions and inaccuracies may be detrimental for future learning;
- (b) The method of assessment of prior knowledge strongly influences the outcomes of learning. Objective assessment methods are connected with positive outcomes. Not or less objective assessment methods such as familiarity ratings and self-estimations do not result in positive outcomes, but are useful to find explanations for effects of prior knowledge on performance.

The general conclusion of their review is that prior knowledge is indeed an effective aid for learning new knowledge. It is suggested that students' reflection on their prior knowledge by assessment may have a facilitating effect on their learning. These conclusions give support for the idea of activation by assessment of prior knowledge as a didactical intervention at the beginning of a new cycle in the learning process.

The effect of assessment is also known from test methodology as an unwanted side effect (Shadish, Cook, & Campbell, 2002). Two aspects are reported:

- (a) The testing-effect occurs when the pre-test is used as a post-test and hence is taken for the second time. It is considered as a threat to the internal validity of the experiment.
- (b) The interaction between the pre-test and the treatment (see also Lana, 1959, 1960, 1969).

Lana and King (1960) analysed the nature of this pre-test sensitization and pointed at similar learning factors as Dochy et al. (1999).

Wilson & Putnam (1982) selected from 32 studies in a meta-analysis concerning pre-test sensitization effects 132 results out of 164 in which randomized groups were used. A pre-test effect was found that cannot be safely ignored: an average effect size of + 0.22 (range between -0.55 and +4.06) was found, with a strong influence of type of outcome, age, and time between pre- and post-test. The effect does not appear to be uniform across the psychological domains. 81 % of the cognitive effects were positive. In other domains (affective, attitude, personality) this fraction was much smaller. Cognitive gains (average effect size = +0.43) are the largest with memory and practice effects when pre- and post-test are the same (see the abovementioned test effect). The studies reported were not considered exhaustive enough to provide definitive statements about conditions for variation of pre-test sensitization.

Next to assessment of prior knowledge, Strangman, Hall & Meyer (2004) concentrate on methods that activate prior knowledge or background knowledge and its impact on reading comprehension and/or recall of curriculum subject areas including science, social studies, and reading. Strangman et al. (2004) make a difference between building up new knowledge and activating existing knowledge. Since the former includes the latter both forms will be discussed here. In order to build up new knowledge, three educational strategies are distinguished: direct instruction, field experience in authentic situations, and text previewing. Direct instruction in order to build up relevant background knowledge led to significantly greater performance on comprehension questions. An alternative for direct instruction is found in acquiring experience in authentic situations. Little evidence is found, though some indications of effects are found. We give some examples from the research literature:

Roschelle (1995) states that the mere confrontation of learners with authentic objects and situations in such an institute as a museum does not build up knowledge by itself (Roschelle, 1995). The Jason-online project, also based on discovery learning, offers digital labs, video interviews of scientists and other multimedia resources. Chatting with scientists doing on-site Arctic research makes it possible to ask questions and building up relevant knowledge in authentic situations (*Jason-online project*). Though learning gains have been reported (Goldenberg, Ba, Heinze, & Hess, 2003), from the evaluation of Hansteen-Izora, Tobin, & Yang (2006) follows that sheer invoking authentic situations is not a guarantee for building up proper knowledge: video recordings and websites facilitated the build up of knowledge more than interactive components. The use of

all sorts of previewing of a text to be learned appears to be effective for understanding and recall. Examples are stating the essence of a text, giving a summary, explaining the text structure, etc. A very explicit manual for reading biology texts is given by Super Read! (*Super read! Strategies for effective reading in biology*).

Based on a considerable amount of research indicating that activation of prior knowledge improves the learners' comprehension and recall of text, Strangman et al. (2004) review six instructional approaches:

- (a) Reflection and recording is a simple straightforward but proven beneficial method of the activation of prior knowledge. Students are prompted to state, write down or (concept-) map what they know about some topic. This approach can be applied also in reciprocal teaching setting.
- (b) Interactive discussion is a more specific form of an instructional approach (a) in which the teacher can direct and all participants can stimulate the discussion. Brain storming and working together on a semantic map can be used as specific variants. The robustness of interactive approaches is not always very impressive, and consistently solid evidence to support the use of an interactive approach to prior knowledge activation is lacking.
- (c) Explanatory answering, especially construction of answers for questions about to-be-learned material has proven to facilitate learning of the material. Pre-testing can be considered as a form of this process (Pressley et al., 1992).
- (d) In the KWL-strategy the different activities can be summarized as: accessing what I **K**now, determining what I **W**ant to find out, recalling what I did **L**earn.
- (e) Computer-assisted activation of prior knowledge with comparing and contrasting new and existing knowledge as key component appear to be successful.
- (f) Interpretation of topic-related pictures can significantly improve reading comprehension for both pictures and text.

Concept mapping is a special form: students are asked to make individually or in small groups with or without computer a graphical representation of a relevant knowledge network. The map can make the link up between new and existing knowledge. Concept mapping is reported to foster conceptual learning, critical thinking, analysis, synthesis, and the development of shared meaning (Daley, Cañas, & Stark-Schweitzer, 2007). Concept maps are used both as a teaching and assessment instrument (Kommers, 1997) and applied in a wide range of disciplines e.g. earth sciences (Rebich & Gautier, 2005), biology (*Biology lessons*). Next to the obvious benefits De Simone (2007) presents limitations: students may find it difficult, time consuming or non-essential. In search of teaching strategies to promote meaningful learning instead of relying on traditional methods that promote rote memorisation (De Simone, 2007). Clayton (2006) surveyed research on application of concept mapping in the training of medical staff and concluded that concept mapping appeared to be promising, but in this particular field the evidence was not sufficient enough for drawing generalizations, since "there is a lack of between-group studies and pre-test to post-test research on knowledge acquisition. Sample sizes are small, there is a lack of instrument reliability and validity, and a lack of control for extraneous variables" (Clayton, 2006).

Strangman et al. (2004) conclude in their review of research results, that the most effective and efficient ways of activation of prior knowledge is direct instruction, student reflection on background knowledge by making it explicit, and activation by asking questions. As has been stated already most research concerned comprehension and recall of texts in science and social studies. Promising and worth further study is a computer supported approach.

We would like to endorse this view taking into account the problem of the reduced presence and availability of the teacher that impels the search for effective and efficient instructional means. Five out of six activation strategies presented by Strangman et al. (2004) require lots of teacher time. Taking also in account the findings of Dochy et al. (1999) activation of prior knowledge by pre-testing seems favourable, while deployment of ICT can minimize teacher overload. In an automated environment the use of closed questions is the most obvious especially when immediate feedback is to be given, but from the study of Strangman et al. (2004) it seems also plausible that the students' making the prerequisite knowledge explicit may contribute to activation effectiveness. A student makes knowledge explicit e.g. by formulating an answer to a question that is

intrinsic to open questioning. However, answers to open questions are hard to handle automatically.

This leads to the question whether effectiveness of pre-test sensitizing also depends on the type of question in the pre-test. Taking into account Dochy et al. (1999) and Strangman et al. (2004) we will focus on two specific types: closed multiple choice questions and (open) short-answer questions.

3 Research questions.

The general research question is, whether an interactive digital educational system with (a) a component '*sensitizing*' by means of a pre-test and (b) a component '*making concepts operational*' by means of interactive assignments and immediate feedback, leads to a better acquirement of scientific (beta) concepts as measured by testing. Next, we wonder, what type of pre-test in this educational system is more effective: a pre-test with short-answer questions (SAQ), or one with multiple-choice questions (MCQ). In short the research questions are:

1. Does an interactive digital educational system with a sensitizing pre-test component and a component for making concepts operational lead to a higher learning gain?
2. Is there a difference in learning gain between a sensitizing pre-test that consist of short-answer questions (SAQ) or of multiple-choice questions (MCQ)?

4 Method.

We successively will discuss the design, the participants, the instruments, the material, the procedure, the scoring, the statistical analysis and the gain estimation.

Design.

Two sets of two groups each are formed by a randomization procedure. The first set consists of a control and a test group that is each given a pre-test, the treatment, and the post-test. The second set consists of a control and a test group that is each given the treatment, and the post-test, but no pre-test. This design has two advantages: (1) The pre-test given to the first set gives an indication of the degree of equivalence of the control and test groups after the randomisation, in addition to allowing measurement of the pre- to post-test gain. If there is a statistical significant difference between the average pre-test scores of the two pre-test groups the whole experiment may be flawed.

(2) Because the first set of control and test groups is given a pre-test and the other set is not given a pre-test, this design makes the pre-test sensitization visible. The pre-test is different from the post-test in our research in order to avoid the testing effect; we aim at the pre-test interaction with the treatment.

In our experiment two different kinds of pre-tests are used, so an extended Solomon Four Group Design (Campbell & Stanley, 1963) is used (see table 1).

Participants.

74 students (year 4 of a six year pre-university school, in Dutch 4VWO), average age 15.5 years, participated in the experiment immediately after summer holidays.

In order to check the equivalence of the at random formed groups, average study results in the preceding year over all subjects were used. No significant differences were found between the six groups using a one-way ANOVA ($F(5,68) = 1.65$ $p = 0.16$).

Because of the great experimental importance of groups 5 and 6, these groups were mixed and divided into two in a second two step-computerised randomisation: at random a student was chosen from the combined group. Subsequently the nearest neighbour was sought on the base of the criteria gender and average school results. The first student was placed at random in one of the two groups, and the nearest neighbour in the other group. This procedure was repeated until all students were assigned to one of the two groups.

One student forgot to make the post-test. His data were removed from the dataset. During the experiment the groups were separated from each other in order to avoid contacts.

Table 1: Extended Solomon Four Group Design

Group	With/without pre-test With/without treatment (R) = Randomisation	Pre-test	Treatment	Post-test
1	without pre-test without treatment (R)			O ₂
2	pre-test O _{1A} without treatment (R)	O _{1A}		O ₂
3	pre-test O _{1B} without treatment (R)	O _{1B}		O ₂
4	without pre-test treatment only (R)		X	O ₂
5	with pre-test O _{1A} and treatment (R)	O _{1A}	X	O ₂
6	with pre-test O _{1B} and treatment (R)	O _{1B}	X	O ₂

Instruments.

Two pre-tests to be administered digitally were made with the commercially available authoring system *Wintoets*. The two pre-tests (O_{1A} en O_{1B}) consisted of 16 short-answer questions (SAQ) and 15 multiple-choice questions. Each subject matter element was present in both a SAQ and a MCQ. The effect of internal diffusion/transfer was taken into account: if some measure activates some subject matter knowledge, strongly associated knowledge is influenced too (Lawson & Chinnappan, 2000). This created the research-technical need to gauge effects at divergent subject matter in order to reduce the threat by diffusion/transfer. Since the *Wintoets* authoring system has extended digital presentation facilities and is able to record necessary research data, it was decided to make also the learning material (= treatment X) as well as the post-test (O₂) on the same platform. An additional advantage was that the student was confronted with only one interface.

The post-test O₂ was made up of 32 questions: 24 questions requiring an answer of one or a few words, six fill-in-the-blank questions and two questions requiring a numerical answer.

By choosing a post-test with open questions we took into account that the gambling element connected with multiple-choice questions causes higher error variances and lower precision (Zimmerman, 2003). Especially polytomous graded open questions appear to be more reliable than multiple-choice questions, but unequivocal a priori statements on validity differences are hard to give, since the domain and purpose of the testing have great influence (Kuhlemeier, Steentjes, & Kleintjes, 2003). On the one side, open questions are usually more difficult than multiple-choice questions: answering an open question requires construction of an answer, lacks the possibility of back reference via the choice items, and the chance on giving a correct answer by elimination of implausible answers is absent. On the other side, from the perspective of sensitization open questions could be an advantage, because of the need to construct the answer. Finally, no immediate feedback was given at either test (O_{1A} O_{1B} O₂).

Intervention ("X")

The educational target of the intervention was a first orientation on the subject matter of various curricular activities at the beginning of the 4th year of the secondary school:

- Chemistry: atomic theory, molecular structure and a part of organic chemistry especially on behalf of biology.
- A lecture of a professor in theoretical physics on structure of matter combined with poster presentations by the students on the same subject matter (Bais, 2004).
- A multidisciplinary project on nanotechnology General science/Physics/Chemistry and English, "*Oscillating cantilevers*" (Ilic & Craighead, 2004) and German "*Nukleare Mikroorganismen*" (Schroeder, 2004).
- Computer science: colour-coding systems.

It was a challenge to link these very diverse subjects into one logical aggregate.

Actually the material ("X") comprised a number of small computer assignments in order to learn new science concepts. Several graphical representations were shown and assignments given in the computer program. The system responded immediately on answers for assignments with concise feedback.

The orienting and sensitizing function of the treatment and the pre-test in the school curriculum are depicted in figure 1.

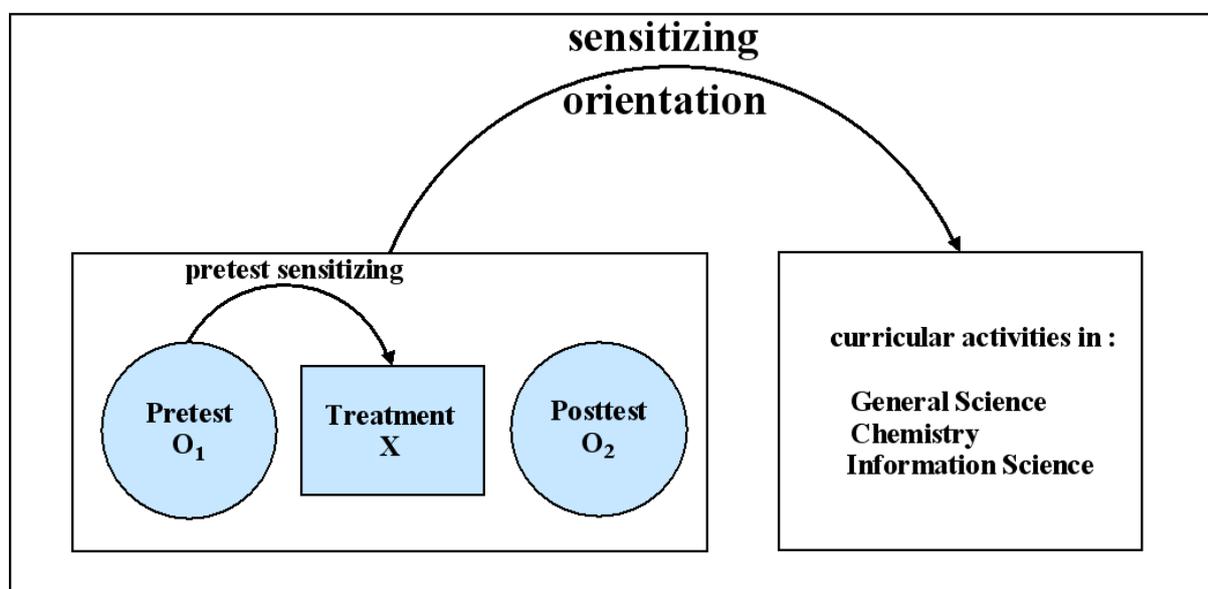


Figure 1: Position of the experiment within the curriculum since the function of the treatment is to sensitize and orientate; the pre-test effect is a nested sensitization.

More specific the elements of the treatment were:

- Use/application of a science data book (*BINAS, 2004*).
- Explication and operationalisation of knowledge of elementary particles and forces.
- Use of powers of ten and logarithmic plots.
- The most abundant elements in the human body (HOCN) and trace elements
- Introduction of conventional colour schemes in molecular modelling.
- Colour generation on computer screens.

After an information screen explaining the purpose of the intervention to the students, it was stated in the next screen (paraphrasing a popular Dutch expression) "that it is not possible to make an elephant from a mosquito"; however, at an atomic level the components appear to be quite the same. After explanation of the concept of *order of magnitude* students were asked to compare mosquito and elephant masses by means of a table with agreed prefixes of units with multiples in powers of 10 (*BINAS, 2004, table 2*). If desired, the use of the Graphical Calculator was explained.

Next to this it was asked explicitly to use the index to find information on structure of matter (BINAS, 2004, table 26). Since the central figure in this table starts with a metal cube, the concept of *molecule* is not shown. The student is asked to state this missing concept (see Figure 2).

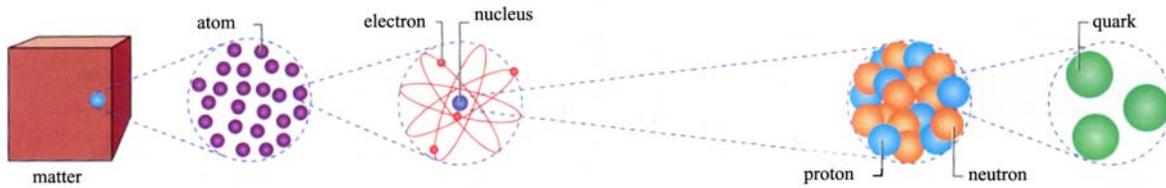
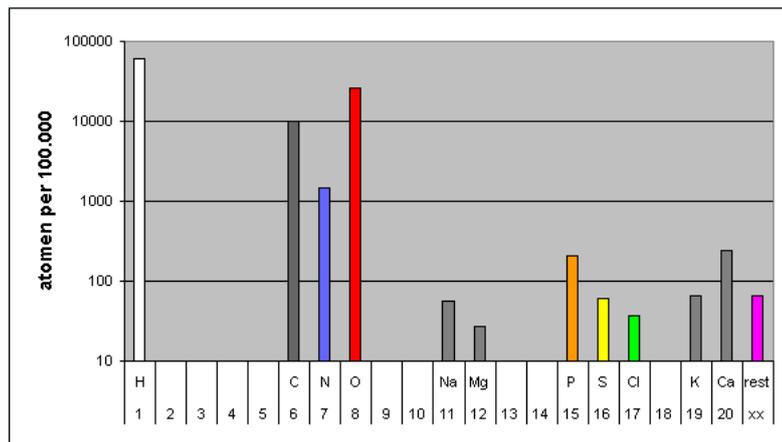
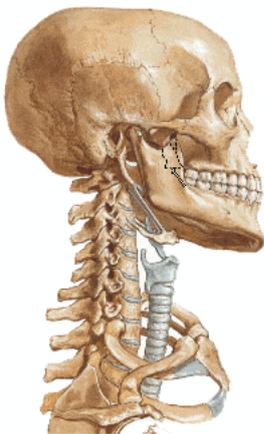


Figure 2: An example of an assignment for activating a science concept. The assignment is: *The cube shown is a piece of metal, which doesn't contain separate [right answer = molecules]* (Source: BINAS, 2004, table 26).

The next assignment was to give a translation of the Greek word $\alpha\rho\omicron\mu\omicron\varsigma$ using BINAS, 2004, table 2 (the Greek alphabet). As extra information the 19th century origin of the specific use of the word was given. Via simple to be answered questions by using the appropriate BINAS-table 26 the attention was focussed to "already known" concepts such as *hadrons, leptons, exchange particles and elementary forces*. An animation of a quark interaction with a gluon was also shown in the treatment of this table. As a form of verbalisation the students were asked to draw a schematic map on a photocopy of this table. In line with the concept *order of magnitude* is the *logarithmic axis*. After explanation of the principle of an axis like that, the student was asked to pick the log axis from three different axes shown (a linear, a logarithmic and a fantasy axis). The use of this type of display was demonstrated by showing the dimensions of a proton, atom, bacterium, mosquito and a human on one axis. In this way it was easy to get an idea of the usual structure size in nanotechnology. Via the theme *oscillating cantilevers* - devices that make it possible to gauge the masse of a few thousand atoms - the focus was set on extreme large numbers and small dimensions.

Figure 3 shows the most important atom types (elements) in the human body as shown by table 34 in BINAS (2004) (*composition of earth, body etc.*).



Het skelet bestaat uit water, eiwit en een vorm van calciumfosfaat. Dit wijst naast calcium, C, H, O en N op het element (geef het symbool :)

Tip: kijk ook in het histogram.

Figure 3: Example of a fill-in question. Note that the ordinate of the histogram is logarithmic and the bars have CPK-colours. Sources: skeleton, courtesy of Ciba-Geigy; the Histogram is free after (BINAS, 2004).

In this histogram a logarithmic ordinate (y-axis) is used. The same data were also shown using a pie diagram. The students were asked to name the most frequent elements in the human body using the *CPK-colours*. Figure 3 shows how the element P in table 34 was put in the spotlight. This section was closed by an explanation of the concept of *trace elements* with a reading text on anaemia and iron and cobalt deficiencies (and of course with simple questions on the subject).

Table 40 in BINAS (2004) (*Elements*) was used to convert atomic size data in Pico meters (pm) into screen representations of filled circles with radius in pixels. The use of the standard graphic editor was explained within this framework. It was also made clear how to use *web safe CPK-colours* in molecular modelling, and the use of *RGB-screen colour codes*.

Small pieces of information were given or pointed at in the treatment. Next, it was asked to apply the information, and based on the answer immediate feedback was given. Obvious bridges linked rather diverse subjects to each other, combining it to one continuous entity. The BINAS (2004) data book played a central role for realising a continuous entity. This book is an important source of information in secondary science education that can be used at all times, including at the official exams. The subject matter was strongly connected to the subjects to be taught/learned in the weeks after the treatment.

The treatment comprised 12 information screens, 13 open questions, 4 fill-in-the-blank questions, 6 multiple-choice questions and 2 true/false questions.

Appendix 1 gives an impression of the science concepts that were dealt with in the tests and the assignments in the interactive digital system.

General procedure

The students were informed, that they were to participate in an educational experiment. The subject matter was connected to the lessons in General Science, Chemistry and Computer Science that would be delivered in the next weeks. The participation would have no negative consequences or whatsoever (on the contrary). Taking the pre-test (O_{1A} / O_{1B}) took 10 minutes, completing the assignments ("X") 40 minutes. The post-test (O_2) took 10 minutes. There were no breaks between the parts.

Scoring procedure.

The computer scored the multiple-choice questions. All answers to open questions were stored in the format (question-ID, student-ID, answer) in a relational database. Using a strict answer protocol the open answers were scored by two independent correctors. In only 1 % of the cases a discrepancy between the two correctors was found. In this case the average score of the two judges was taken.

Statistical Analysis.

An analysis of variance and a multiple comparison according to Bonferroni (significance level 5%) had been executed with SPSS. A two-way ANOVA analysis was performed with the VISTA 6 statistical package. A test-item analysis was done with the TIAPLUS program version 2.1. Cronbach's alphas were calculated for all tests (O_{1A} , O_{1B} , and O_2). The alphas for the pre-tests were separately calculated for the multiple-choice parts and the short-answer parts.

Calculating learning gain.

It is not possible to calculate learning gains if post-test results only are taken into account, since pre-treatment levels have to be known also. Two problems have to be dealt with: the sensitizing pre-test, and the correction for individual or group pre-treatment levels. The Solomon Four Group Design is a solution for the first problem as has been explained above. A pre-test corrected learning gain calculation has been devised for solving the second problem.

In several test-retest experiments concerning the school subjects French, Computer Science, and Chemistry a strong empirical relationship between pre and post-test was found. This relation can be used in order to eliminate the variable 'pre-test' and to calculate learning gain (Bos, Terlouw, & Pilot, 2007b). An explanation:

If pre-test scores are divided by the maximum pre-test score, and this variable is called x , ($x = \text{pre-test score}/\text{maximum pre-test score}$ $0 \leq x \leq 1$) and the same is done with the post-test scores

and the maximum post-test score, and this variable is called y , the growth factor $f = y/x$ can be described with the power function $f = x^{-B}$. The exponent B is a robust measure of the learning gain in pre-test-treatment-post-test designs ('OXO'-designs). Normally the post-test score is larger than the pre-test score (otherwise nothing seems to be learned); therefore the exponent B is between 0 and 1. Statements of statistical significance of differences between learning gains can be supported using estimations of the error in the parameter B . A nominal categorization of the knowledge growth exponent B is depicted in table 2 which is based on a calibration with data from a review of Hake (1998).

Table 2: Nominal scale for the knowledge growth exponent B

Exponent	Gain characterization
$B \leq 0.40$	Low
$0.40 < B < 0.60$	Average
$B \geq 0.60$	High

A special problem gives the gain calculation with group 4. Formally gain calculation is impossible since the group does not make a pre-test, but on the base of the equivalence of the six groups an estimation of the group gain B may be calculated from

$$B = -\log(\langle y \rangle / \langle x \rangle) / \log(\langle x \rangle)$$

The angle brackets $\langle \dots \rangle$ signify group averages and scores are normalized so that $0 < \langle y \rangle < 1$ and $0 < \langle x \rangle < 1$. This method using group averages (1) may yield lower B values than when individual student scores are used, and (2) there is no information on the B parameter error.

As an extra control also the classical effect size categories according to Cohen (1988) is reported. Cohen (1988) suggested that as a very rough rule of thumb $d = 0.2, 0.5,$ and 0.8 imply respectively "small," "medium," and "large" effects. Effect sizes of more than 3 standard deviations calculated with Cohen's method are considered as extreme. At this point it should be stressed again, that effect size is not the same as learning gain.

Test analysis.

A correction for guessing of $-1/(k-1)$ was applied to the pre-test scores of multiple-choice questions, with $k=4$ for four choice questions and $k=2$ for true/false questions. Averages and standard deviations after correction for guessing are given in table 3.

Table 3: Pre-test O_{1A} and O_{1B} scores after correction for guessing (maximum = 100)

Group	Pre-test	After correction for guessing		Students
		Average	Std.dev.	N
2	O_{1A}	16.9	12.1	8
3	O_{1B}	15.8	10	8
5	O_{1A}	21.8	7.31	13
6	O_{1B}	16.7	10.7	14
All		18.1	9.9	43

ANOVA of pre-test results did not show a statistical significant difference between the four groups before correction for guessing $F(3,39) = 0.580$ ($p=0.63$) nor after correction for guessing $F(3,39) = 0.883$ ($p=0.46$). The groups can be considered equivalent before the intervention. The values of the pre-test corrected for guessing were used for the gain calculation.

Very low Cronbach alpha coefficients were calculated for the pre-tests. In a computer simulation the alphas did not differ significantly from zero. Cronbach's alpha did differ significantly from zero for the parts with short-answer questions (in table 4 indicated with an *). In table 4 Cronbach's alpha is also given for the post-test that completely consisted of short-answer questions.

Table 4: Cronbachs coefficient alpha and significance (P).

	α	Number of students	Number of questions	p
O _{1A} SAQ + MC (total)	0.11	21	31	0.42
O _{1A} SAQ-part	0.49*	21	16	0.023
O _{1A} MC-part	-0.41	21	15	0.88
O _{1B} SAQ + MC (total)	0.23	22	31	0.25
O _{1B} SAQ-part	0.48*	22	16	0.025
O _{1B} MC-part	-0.16	22	15	0.74
O ₂ SAQ	0.93*	74	32	0.000

In table 5 some results of the test item analysis by the TIAPLUS program of CITO (Central Dutch Test Institute) of the post-test O₂ are given. Except for question 17 all p-values (average percentage score per question) were between 10 and 90. No question had a negative item-rest correlation. No question would have a significant effect on Cronbach's alpha. Consequently all questions were used in the calculations.

Table 5: Global parameters of the test-item analysis of post-test O₂

Number of test persons	74	Number of questions	32
Average p-value	41.3	Std. error	4.3
Coefficient Alpha	0.93	Std error Coeff. Alpha	0.01
90% - interval Alpha	0.91-0.95		
Average item/rest correlation	0.59		

Calibration of the tests.

Since the low Cronbach alphas of the pre-tests O_{1A} and O_{1B} could evoke some doubt about the internal consistency, it seemed necessary to calibrate the tests in a separate experiment. Each of the 32 post-test questions is connected to a different but strongly related question in the pre-tests. All questions from O_{1A}, O_{1B} and O₂ were gathered in one file, but clustered in the 32 pieces of subject matter. To 70 randomly chosen students from upper level secondary education (comparable to the 74 students in the study) 32 questions were offered, while the question was at random chosen from O_{1A} or O_{1B} or O₂. Each student was offered a 32-item test on 32 elements of subject matter, but each test was in fact different because of the random choice. The scoring procedure was equal to the one in the main experiment, using the same correction for guessing. Subsequently averages scores were calculated per student, but grouped by question source (O_{1A}, O_{1B}, O₂). Tests on the same subject matter are equivalent when (*ceteris paribus*):

- (a) There is no difference between mean scores.
- (b) A high linear correlation between the outcomes exists.

The results are given in table 6.

Table 6: Average scores and standard deviations grouped for questions from O_{1A}, O_{1B} en O₂ and total test scores. Scores are given on a 0-100 scale. (N = 70).

Source	Average	StDev
O _{1A}	52.7	32.0
O _{1B}	48.0	33.7
O ₂	50.9	31.0
Total	50.5	31.9

An analysis of the average scores revealed that

- (a) No statistical difference existed between the averages using an ANOVA analysis $F(3,276) = 0.263$ ($p = 0.852$), and

(b) The intra class correlation coefficient of the results per student grouped by test source was $ICC(3,1) = 0.875$ (model: two way mixed, single measure).

Because there are no differences in average scores and there is a high correlation between the three tests, it can be concluded that the tests O_{1A} , O_{1B} en O_2 are equivalent.

5 Results

Research question 1: Does an interactive digital educational system with a sensitizing pre-test component and a component for making concepts operational lead to a higher learning gain?

The average post-test scores for students in the six groups are shown in table 7.

Table 7: Post-test O_2 Results (maximum = 100)

Group	Design	Average	Std. Dev.	N
1	O_2	10.9	6.1	17
2	$O_{1A} O_2$	20.8	5.8	8
3	$O_{1B} O_2$	22.2	9.9	8
4	$X O_2$	51.8	15.8	15
5	$O_{1A} X O_2$	66.8	14.1	12
6	$O_{1B} X O_2$	67.6	9.2	14
	Total	41.3	26.0	74

An analysis of variance with ANOVA of the post-test O_2 results gave a significant difference between the 6 groups: $F(5,68) = 67.47$ $p = 5.24 \cdot 10^{-25}$.

A Bonferroni multiple comparison between the groups is depicted in table 8:

Table 8: Significance (p) of post-test score differences based on Bonferroni multiple comparisons between groups.

Group	Design	1	2	3	4	5	6
1	O_2		0.63	0.31	0.000	0.000	0.000
2	$O_{1A} O_2$	0.63		1	0.000	0.000	0.000
3	$O_{1B} O_2$	0.31	1		0.000	0.000	0.000
4	$X O_2$	0.000	0.000	0.000		0.012	0.004
5	$O_{1A} X O_2$	0.000	0.000	0.000	0.012		1
6	$O_{1B} X O_2$	0.000	0.000	0.000	0.004	1	

Our conclusions:

Groups 2 and 3 score higher than group 1, but this difference is not significant.

Groups 5 and 6 score significantly higher than group 2 and 3. The treatment 'making concepts operational', including feedback (= "X") gives a considerable effect.

Groups 5 and 6 score significantly higher than group 4. The combination of the treatment "X" and a pre-test gives a strong effect.

Group 3 score higher than group 2, but this difference is not significant. This indicates equivalence of group as measured with the pre-tests O_{1A} and O_{1B} .

The difference between group 5 and 6 is not significant. The influence of O_{1A} and O_{1B} is equal.

Group 4 scores significantly higher than group 1: the treatment “X” has influence

The learning gain exponent B was calculated from the pre-test and post-test values. Between groups 2 and 3 there was no significant difference ($p=0.22$). For the joint groups 2&3 the exponent $B = 0.10 \pm 0.071$.

Between groups 5 and 6 there was no significant difference ($p=0.11$). For the joint groups 5&6 the exponent $B = 0.79 \pm 0.021$.

Using the average pre-test results of groups 2,3,5 and 6 for group 4 an exponent $B = 0.62$ was calculated. The effect size calculated according to (Cohen, 1988) of the combined groups 5 and 6 compared to the combined groups 2 and 3 was 3.5.

Research question 2: Is there a difference in learning gain between a sensitizing pre-test that consist of short-answer questions (SAQ) or of multiple-choice questions (MCQ)?

The average post-test scores of the ‘sensitizing pre-test + treatment’ groups 5 and 6, grouped to pre-test question type, are depicted in table 9.

Table 9: Post-test scores of the ‘sensitizing pre-test + treatment’ groups 5 and 6 grouped to pre-test question type.

Question type in pre-test	Average	Std. Dev.	N
Short-answer question	69.6	45.5	340
Multiple-choice question	71.7	43.8	338
Total	70.6	44.7	678

An ANOVA analysis revealed no significant differences ($p=. 52$)

6 Conclusions

Research question 1: Does an interactive digital educational system with a sensitizing pre-test component and a component for making concepts operational lead to a higher learning gain?

As indicated above the interactive digital learning system consisted of two parts:

- (a) A component ‘sensitizing pre-test’.
- (b) A component ‘making operational’ of science concepts to-be-learned by means of assignments and appropriate feedback.

Applying this two-component system led to significant higher post-test scores in comparison to the groups in which none or only one of the components was applied. The calculated joint learning gain $B = 0.79$ can be called *very high* applying the nominal scale of table 2. This learning gain is also significantly higher than applying the system without pre-testing. Applying only the component ‘making operational’ also led to significantly higher post-test scores in comparison to the groups with only a ‘sensitizing pre-test’ component or no components at all. The learning gain was also relatively high ($B = 0.62$) taking into account table 2. Applying only the component ‘sensitizing pre-test’, without any follow up by assignments or feedback in the component ‘making operational’ did not have a significant effect.

Research question 2: Is there a difference between sensitizing with short-answer questions (SAQ) or with multiple-choice questions MCQ?

There is no difference between learning results of a pre-test with short-answer questions of with multiple-choice questions.

7 Discussion

Students achieve the highest learning gain (B is about .80) in learning science concepts when a sensitizing pre-test is directly followed by a focused multimedial, interactive system with assign-

ments and direct feedback. A high learning gain is also already achieved without pre-testing (B is about .60). Therefore, from a didactic perspective it has sense to connect pre-testing directly with a teaching strategy that consists of a good explanation, followed by questions and immediate feedback in order to enhance learning. Only applying the pre-test will not result in learning.

Following the review of Strangman et al. (2004; see the theoretical framework) the good results found can be explained from a combination of teaching strategies that are considered as effective: activating prerequisite knowledge by asking questions before (pre-test), and by building up prerequisite knowledge using direct instruction, asking questions and giving feedback.

The type of pre-test-questions – multiple-choice or short-answer questions – does not look to matter according to our research. Moreover, no significant pre-test-effect was also found with two-choice questions in an adjoining experiment, and it is not to be expected that such an effect will be found in a large-scale experiment. However, because the short-answer questions are not so much different from the multiple-choice questions used in this research, it is possible that real open questions will make a difference. From the perspective of educational efficiency there is a problem here: The available off-the-shell-software for the automatic scoring of open questions is still time consuming for teachers, because a lot of control-afterwards of the scoring is necessary. Therefore, taking into account the effectiveness of the instructional strategy applied in our research and the need for educational efficiency, it looks obvious to apply a digital, interactive system with multiple-choice pre-test questions. The last also offers the opportunity - we did not do that in our research for experimental reasons – to give immediate feedback on the answers given. We expect that the application of a combination of pre-test and immediate feedback will lead to significantly higher learner gains than the application of a pre-test alone.

We like to give two helps for the instructional practice:

The first help for the instructional practice is the idea that the design of the experiment could serve as an instructional design for an introductory (science) module. The instructional design consists of a digital learning environment in which a multiple-choice pre-test with immediate feedback is embedded, directly followed by a number of screens with digitally controlled assignments with also immediate feedback. Students can work with such an introductory module before the new course(s) in their own chosen time, pace, and place. Process results of students from this introductory module could be interesting for the teacher at the beginning of the new course to take into account for the teaching. Such an approach is not new: The CAI-package SCOOR (Paulides & Pilot, 1996) – a program meant for detecting and removing deficiencies in the knowledge-base of starting students of Professional Higher Education – is a comparable approach. However, the pre-test in the CAI-package SCOOR has particularly an allocating function –students are allocated to one or more specific modules dependent on the pre-test-score – and not a sensitizing function, but the pre-test could work in this way. A high learning gain ($B = .73$) could be calculated from (still) available pre-test / post-test data of a group of students that followed the SCOOR chemistry module. For the non-SCOOR group $B = .12$ (SCOOR, 1986). We expect even a higher learning gain now, taking into account the increase of multimedia opportunities and asynchronous access of the present digital systems.

As a second help for the instructional practice we like to pose a discussion point: Pre-test-sensitization – may be in combination with other forms for activation and building up prerequisite knowledge (see theoretical framework) – could be helpful for concept development in the context-concept approach in innovative science education in secondary education (Bulte et al., 2005). The smooth execution of tasks within the context chosen implies that relevant conceptual networks are available and transfer of these networks are possible which appear to be a problem (Pilot & Bulte, 2006). Also Strangman et al. (2004) indicate that the mere use of authentic situations does not automatically lead to the development of prerequisite knowledge. Pre-test sensitization could facilitate availability and transfer of existing conceptual networks. Probably, a strategy for activating and building up prerequisite knowledge should be also followed in order to stimulate learning in authentic situations. The results found in this experiment can, may be, play a role in this instructional design.

8 Appendix 1.

Schematic overview of subject matter in the treatment.

Scientific Concepts

10⁶ g

Ceci n'est pas un éléphant

10⁻³ g

atoms per 100,000

10⁻¹⁶g

450 nm

10⁻²⁴g

electron

positively charged sphere

ELEMENTS			
Hydrogen 1	Strontium 38	42	
Lithium 3	Barium 56	56	
Carbon 6	Iron 26	50	
Oxygen 8	Zinc 30	56	
Phosphorus 15	Copper 29	55	
Sulphur 16	Lead 82	50	
Magnesium 12	Silver 47	108	

A part of the CPK colour scheme

C	H	O	N	Cl
204	255	162	61	
184	232	182	209	
204	233	231	211	

Web-safe colours
N=51 N=0..5

matter

atom

electron

nucleus

proton

neutron

quark

9 References

- Anderson, J. R., & Schunn, C. D. (2000). Implications of the ACT-R Learning Theory: No Magic Bullets. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 5). Mahwah, NJ: Erlbaum.
- Bais, S. (2004). *Nooit meer rechtdoor. Over het wankele evenwicht tussen waarneming en verbeelding [Never more straight on. About the unstable balance between perception and imagination]*. Deventer: Thieme.
- BINAS (2004). *Biology lessons*. Retrieved Oktober 31 2006 from <http://www.biologylessons.sdsu.edu/about/index.html>
- Bos, A. B. H., Terlouw, C., & Pilot, A. (2007a). *Het effect van pretest-sensitisatie bij ontdekkend leren met behulp van een simulatie in het voortgezet onderwijs*. Paper presented at the ORD 2007, Zorgvuldig en Veelbelovend Onderwijs, Proceedings van de 34e Onderwijs Research Dagen, Groningen.
- Bos, A. B. H., Terlouw, C., & Pilot, A. (2007b). A Pre-test-Corrected Learning Gain. *in preparation*.
- Bulte, A. Klaassen, K., Westbroek, H., Stolk, M. Prins, G. Genseberger, R., et al. (2005). Modules for a new chemistry curriculum. Research on the meaningful relation between contexts and concepts. In P. Nentwig & D. Waddington (Eds.). *Making it relevant. Context based learning of science* (pp. 273-299). Münster: Waxman.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally College Publishing Company.
- Clayton, L. H. (2006). CONCEPT MAPPING, An Effective, Active, Teaching-Learning Method. *Nursing Education Perspectives*, 27(4), 197-203.
- Cohen, J. (1988). *Statistical power analysis for the behavioural science* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Daley, B. J., Cañas, A. J., & Stark-Schweitzer, T. (2007). CmapTools: Integrating teaching, learning, and evaluation in online courses. *New Directions for Adult and Continuing Education*, 2007(113), 37-37.
- De Simone, C. (2007). Applications of Concept Mapping. *College Teaching*, 55(1), 33-36.
- Dochy, F., Segers, M., & Buehl, M. M. (1999). The relation between assessment practices and outcomes of studies: The case of research on prior knowledge. *Review of Educational Research*, 69(2), 145-186.
- Goldenberg, L., Ba, H., Heinze, J., & Hess, A. (2003). *JASON Multimedia Science Curriculum impact on student learning: Final evaluation report*. New York: Center of Children and Technology, Educational Development Center.
- Hake, R. R. (1998). Interactive-engagement vs. traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *Am. J. Phys.*, 66, 64-74.
- Hansteen-Izora, M., Tobin, B., & Yang, S. (2006). *JASON-project online design review. Learner study*. Retrieved November 3 2006 from http://www.stanford.edu/~btobin/courses/106/jason_online/design_review_site/pages/learner_study/details.htm#conclusion
- Ilic, B., & Craighead, H. G. (2004). Attogram detection using nanoelectromechanical oscillators. *J. Applied Physics*, 95(7), 3694-3703.
- Jason-onlineproject*. Retrieved November 3 2006 from <http://www.jasonproject.org/>
- Kommers, P. (1997). Concept mapping. Special double issue on concept mapping. *Journal of Interactive Learning Research*, 8(3-4), 281-502.
- Kuhlemeier, H., Steentjes, M., & Kleintjes, F. (2003). *De gelijkwaardigheid van open en meerkeuzevragen bij wiskunde. Effecten van vraagtype en scoringswijze op gemeten vaardigheden, betrouwbaarheid, moeilijkheid en afnametijd*. Arnhem.
- Lana, R. E. (1959). Pre-test-treatment interaction effects in attitudinal studies. *Psychological Bulletin*, 56, 293-300.

- Lana, R. E. (1960). A further investigation of the pre-test-treatment interaction effect. *Journal of Applied Psychology*, 43, 421-422.
- Lana, R. E. (1969). Pre-test Sensitization. In Rosenthal & Rosnow (Eds.), *Artifact in Behavioral Research*. New York: Academic Press.
- Lana, R. E., & King, D. (1960). Learning factors as determiners of pre-test sensitization. *Journal of Applied Psychology*, 44, 189-191.
- Lawson, M. J., & Chinnappan, M. (2000). Knowledge Connectedness in Geometry Problem Solving. *Journal for Research in Mathematics Education*, 31(1), 26-43.
- Osborne, J., & Hennessy, S. (2003). *Literature review in Science Education and the role of ICT: promise, problems and future directions* (No. 6). Bristol: NESTA Future Lab.
- Paulides, J. P., & Pilot, A. (1996). SCOR for Windows. In M. van Geloven & A. Pilot (Eds.), *Multimedia in het hoger onderwijs*. Groningen: Wolters-Noordhoff.
- Pilot, A., & Bulte, A. M. W. (2006). The use of "contexts" as a challenge for the chemistry curriculum: its successes & the need for further development and understanding. *International Journal of Science Education*, 28(9), 1087-1112.
- Pressley, M., Wood, E., Woloshyn, V. E., Martin, V., King, A., & Menke, D. (1992). Encouraging Mindful Use of Prior Knowledge: Attempting to Construct Explanatory Answers Facilitates Learning. *Educational Psychologist*, 27(1), 91-109.
- Rebich, S., & Gautier, C. (2005). Concept mapping to reveal prior knowledge and conceptual change in a Mock Summit Course on global climate change. *Journal of Geoscience Education*, 53(4), 355-365.
- Ritzen, J. (2006). Hoger onderwijs tussen kennis en koopje. *TH&MA*(1).
- Roes, T. (2001). *De sociale staat van Nederland*. (Vol. 2001-14). Den Haag: SCP.
- Roschelle, J. (1995). *Learning in interactive environments: prior knowledge and new experience*. Retrieved November 3 2006 from www.exploratorium.edu/IFI/resources/museumeducation.html
- Schroeder, B. (2004). *Nukleare Mikrobatterien [Nuclear micro batteries]*, from <http://www.heise.de/bin/tp/issue/r4/dl-artikel2.cgi?artikelNr=18351&zeilenlaenge=72&mode=html>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.
- Strangman, N., Hall, T., & Meyer, A. (2004). *Background knowledge instruction and the implications for UDL implementation*. Retrieved October 23 2006 from http://www.cast.org/publications/ncac/ncac_backknowledgeudl.html
- Super read! Strategies for effective reading in biology*. November 1 2006 from <http://www.phschool.com/science/biosurf/superread/index.html>
- Tweede_Fase_Adviespunt. (2005). *Zeven jaar Tweede Fase, een balans*, from www.tweedefaseloket.nl
- Willson, V. L., & Putnam, R. R. (1982). A meta-analysis of pre-test sensitization effects in experimental design. *American Educational Research Journal*, 19(2), 249-258.
- Zimmerman, D. W., Williams, R.H. (2003). A New Look at the Influence of Guessing on the Reliability of Multiple-Choice Tests. *Applied Psychological Measurement*, 27(5), 357-371.