Maastricht University

A Dataset of Three Educational Technology Experiments on Differentiation, Formative Testing and Feedback

Document status and date:
Published: 01/01/2015

Document Version:
Publisher's PDF, also known as Version of record

Document license:
Taverne

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Download date: 01 May. 2024

A dataset of three educational technology experiments on differentiation, formative testing and feedback

Carla Haelermans, Joris Ghysels and Fernao Prince

Carla Haelermans is an assistant professor of Top Institute for Evidence-Based Education Research (TIER) at Maastricht University. Joris Ghysels is an assistant professor of Top Institute for Evidence-Based Education Research (TIER) at Maastricht University. Fernao Prince is a biology teacher at Sophianum. Address for correspondence: Dr Carla Haelermans, PO Box 616, Maastricht 6200 MD, the Netherlands. Email: Carla.Haelermans@maastrichtuniversity.nl

Abstract

This paper describes a dataset with data from three individually randomized educational technology experiments on differentiation, formative testing and feedback during one school year for a group of 8th grade students in the Netherlands, using administrative data and the online motivation questionnaire of Boekaerts. The dataset consists of pre- and posttest scores for each experiment, as well as background information on the students, such as gender, age, ability, grade repetition and neighborhood characteristics. It also includes motivation of students. The dataset, or parts of it, could be interesting to use in future studies, focusing on effects of the use of educational technology, for comparison with similar and/or slightly different experiments, and for studies such as meta-analyses or replication studies, especially given the individual experimental setup.

Dataset

Location and doi: ResearchGate, doi: 10.13140/2.1.4407.9049

Creator: Carla Haelermans

Date: October 30, 2014

Format: Stata data file (*.dta)

Restrictions to use (if any): Request permission to use by sending request, including short project description to data owner.

Introduction

It has become increasingly popular to use educational technology in class in order to be able to serve students in a more individual way and to be able to introduce certain didactical aspects without tremendously increasing a teachers' workload (eg, Haelermans, Ghysels & Prince, 2014). It is believed that using educational technology in a smart way may even improve student performance (Banerjee, Cole, Dufflo & Linden, 2007; Burns, Kanive & DeGrande, 2012).

The dataset introduced here contains data regarding three experiments which evaluate the above claims in a realistic setting, namely Biology classes of 8th grade (second year of secondary education in the Netherlands) prevocational students (the lowest track in the Dutch tracking system). One hundred fifteen students (five classes) were individually randomized to an experiment and control group.

In a first experimental period, digital differentiation was offered and the main results of this experiment are reported in Haelermans *et al* (2014). In a second and third period, the same

experimental group was exposed to two types of formative testing: formative testing with extensive, individualized feedback, and simple formative testing with only basic feedback (total score). The evaluation of the latter experiments, reported in Ghysels, Haelermans and Prince (2014), suggests two frameworks for the appreciation of the experiments. Either the second and third experiments are treated as independent, because no signs of transfer are observed, or they are seen as an efficiently cumulative effort which trains students more thoroughly in a first period, after which a simpler setup of formative testing is sufficient to yield learning benefits.

Further analyses of the data may reveal more complex cumulative patterns, however, or may shed light on motivation aspects that have not been explored yet.

The dataset contains a treatment dummy, as well as pre- and posttest scores for all students on each of the experiments, aggregated responses to the online motivation questionnaire of Boekaerts (2002), at three moments during the school year (after a period of fully digital classes, fully traditional classes and combined classes), and students' preferences for each of these three types of classes.

The background variables in the dataset are gender, age, ability, grade repetition, a dummy for new intake students, class, postal codes and place of residency, and, based on the latter, the average income of the neighborhood.

Details on the randomization process are given in the first part of the research method section; the timeline and content of the three experiments are described in the second part, while the third part concludes by outlining the structure and data sources of the developed dataset.

Research method
Randomization process
Randomization took place at the beginning of the school year 2012/2013. Stratified randomization at the student level was applied, to make sure that the experiment and control group were equally distributed regarding the students' primary school ability test score, the class they belong to, their gender and age, and state of grade repetition. In effect, a simple stratification algorithm was implemented, allocating students alternatingly to the intervention or the control group after a random start using a list of students sorted on the previously indicated characteristics within each of the five participating classes. From the total of 115 students, 57 students were assigned to the treatment group and 58 to the control group. All three experiments had the same treated and control students.

Content and timeline of the experiments
In the first experiment, all students wrote small weekly formative tests that covered the material of the week before (about 10 minutes, on the computer). The treatment consisted of a differentiated learning process, with content levels adjusted weekly to the student's performance on the small weekly test, whereas the control condition was a constant level, in accordance with the track of the student, maintained throughout the experimental period. The levels offered were modeled according to three different tracks in Dutch secondary education: the practical prevocational track, the theoretical prevocational track and the higher general track. Note that it was possible for treatment group students to follow a different track each week, depending on their score on the test. The lessons at the practical prevocational track level were written in more simple language, using fewer words and less complicated sentences. Furthermore, the pace of the exercises would be a bit lower, which means that these students studied the minimum amount of topics. The content of the basic topics did not differ between the groups. The theoretical prevocational track had some extra topics compared with the practical track, and more difficult explanation and more challenging exercises. This was also the case for the higher general track,

where more topics were discussed (in most cases slightly off topic, but interesting to the students). Control students were always offered learning content on the intermediate level, the theoretical track.

In the second experiment, all students wrote small weekly formative tests, and the students in the treatment group got specific feedback, depending on their answer, after each question. Feedback was provided on each question answered by students. The feedback focused on whether and why a certain answer was wrong, if the question was answered incorrectly; but feedback was also provided to repeat why this was indeed the correct answer, if the question was answered correctly. At the end of the formative test, treatment group students were provided with an overview of which questions they answered correctly and which questions they got wrong. Control group students did not get specific feedback on their answers during the test, and they were only able to see which questions they answered correctly and which questions they got wrong at the end of the course.

In the third experiment, the treatment group was getting the small weekly formative tests, whereas the control group did *not* get these tests. At the end of this weekly test, treatment students were provided with an overview of which questions they answered correctly and which questions they answered wrong, without further feedback. In the meantime, the control group students worked in their own digitalized environment on the topic of that week and/or made (digital) homework assignments.

Summarizing, the research questions of the three experiments can be phrased as:

1 Is week-for-week differentiation of digital learning content into three levels based on a short multiple choice test beneficial for Biology learning among 13-year-olds?
2 Does the provision of extensive, individualized feedback following immediately after a short, multiple choice, formative test yield an additional learning benefit?
3 Is a short, multiple choice, formative test with basic feedback beneficial for Biology learning?

Figure 1 shows the timeline of the school year. It shows the timing of the three experiments in the school year, of which the first experiment consists of three periods, each with their own pre- and posttests, as well as the moments at which the motivation questionnaires were filled out. During the school year, the students had two lessons (50 minutes per lesson) of Biology each week.

During the first experiment, both these classes were computerized, and all students used the digital instruction material, which was a combination of the digital instruction package from the book publisher, sections from the book that were digitalized by the teacher and additional digital material from the Internet (such as exercises, puzzles, short movies, etc).

During the two other experiments, one of the weekly lessons relied on computer-driven instruction and took place in a computer room, whereas the other lesson was teacher directed and took place in a regular classroom, with some group work from time to time. All students used digital instruction and assignments, as well as their book, together with classroom instruction, to study the course content. The digital instruction consists of the same combination of materials as in the first experiment.

For each of the three experiments, the period started with a pretest in the first class of the first week, and ended with a posttest, in the last class of the last week. Furthermore, every (first) computerized class of the week, the small test was written (depending on the experiment only by treatment students, or by all students).

The first experiment consisted of 12 weeks in total, consisting of three topics of each 4 weeks. Note that the treatment, which will be discussed later, only affected the content that is discussed, not the skills needed to process the content or other aspects that could influence the pretest of the next topic, because of changed study behavior. The first topic was "metabolism and respiration,"
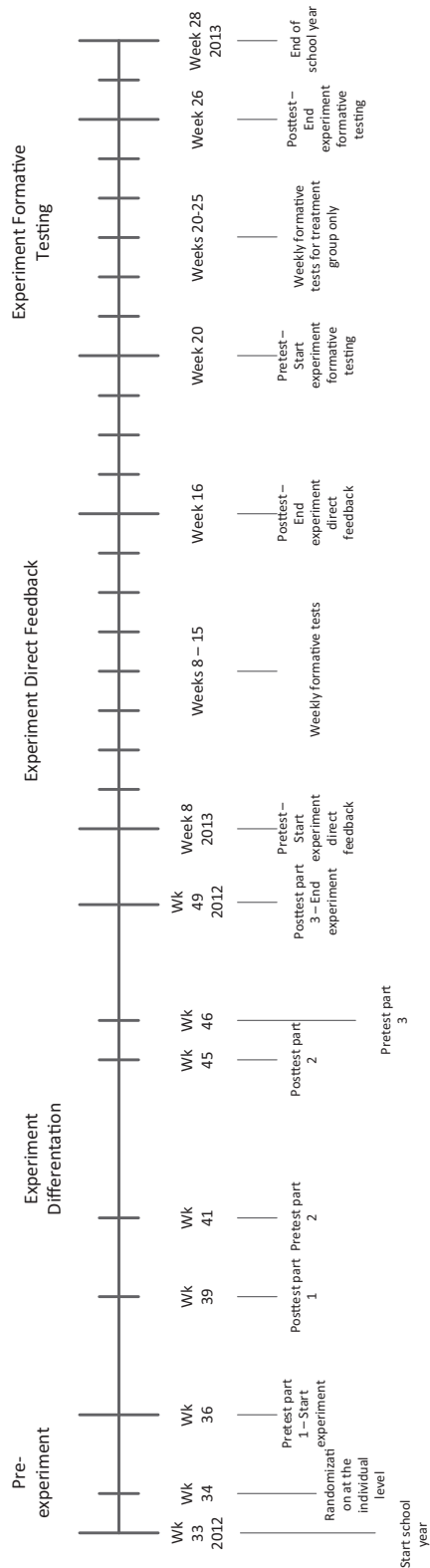
Figure 1: Overview of the timeline of school year

the second topic was "blood circulation" and the third topic was "your health." Because the school wanted students to be graded separately for all topics, we have a pre- and posttest for each of the three topics. The first pretest took place in week 36 and contained easy questions on the topic of the first 4 weeks. In week 39, the first posttest was written. The second pretest, on the second topic, was written in week 41, with again 4 weeks of treatment and a posttest in week 45 (note the second period included 1 week of holiday). The third pretest was written in week 46 and the third posttest was written in week 49, which concluded the experiment.

The second experiment started in week 8 and ended in week 16 of 2013. The topic of this second experiment was "sexuality and relationships." The third experiment, on information provision via weekly formative tests, consisted of 6 weeks in total, from week 20 to week 25 of 2013. The topic of the course taught during these 6 weeks was "heredity and evolution." In between the second and third experiments, there were a couple of holiday weeks and a lesson-free week.

It should be noted that the cumulative nature of the experimental process requires careful attention in any subsequent analysis. Students were exposed to three experiments in a row and hence it cannot be excluded that some transfer may have occurred. However, none of the statistical analyses conducted so far (Ghysels *et al*, 2014; Haelermans *et al*, 2014) have revealed any sign of transfer. At the starting point of any experiment, the pretest scores of the experiment and control groups were very similar. The latter suggests that the benefit of the experiments was closely linked to the topic and/or wore out quickly. Yet future research may look into the potentially cumulative nature of the observed benefits in the second and third experiments more thoroughly.

Structure and data sources
Pre- and posttests
The pre- and posttest consisted of questions on the covered topic that come from a test data bank with questions that have been used for years by this school for Biology. Test items are therefore not scientifically validated, but are selected on face validity by the various Biology teachers which have been involved in the course in recent years.

All tests are computerized, which facilitates the subsequent use of test results in the experiment. The pretest was a multiple choice test and took about 15 minutes. It is meant to reflect prior knowledge students may have on the topic to be introduced.

The small weekly (formative) tests also consisted of only multiple choice questions, lasting for about 10 minutes. The tests check on topical knowledge students should have acquired in the two previous classes and the corresponding homework.

Finally, the posttest (of every experiment) was written during one Biology class, and lasted for about 50 minutes. Different from all others tests, the posttests included both multiple choice questions (about 75% of the test) and open questions (about 25% of the test). The tests probe for knowledge and understanding about the topic of the teaching period the test is closing.

All pre- and posttests, as well as the formative tests, are scored between 0 and 100, representing the percentage of questions answered correctly.

Motivation questionnaires
Motivation is measured based on the validated online motivation questionnaire developed by Boekaerts. Boekaerts (2002) provides an extensive discussion of the associated literature and a theoretical framework on which the the questionnaire is based. This questionnaire consists of 25 questions that have to be answered based on a 4-point Likert scale, ranging from *not agree at all* (1) to *very much agree* (4). The answers to these questions can be reduced to six components, namely: self-efficacy (SEFF), success expectancy (SE), task attraction (ATTR), perceived utility (PU), task

Table 1: Descriptive statistics on background characteristics

	Obs.	Average	St. Dev.	Min	Max
Age	115	13.20	0.52	12	15
Female	115	0.49	0.50	0	1
Grade repetition	115	0.02	0.13	0	1
New intake students	115	0.02	0.13	0	1
Primary school ability test total score	111	530.65	5.74	516	545
Average monthly income neighborhood (€)	115	2107.12	319.34	1532	3358

anxiety (TA) and intended effort (IE). The score on these components is an average of the answers to the underlying questions, where answers to negatively stated questions have been reversed.

Student background characteristics

Data on student background were received from the school administration of the participating school. Table 1 presents the descriptive statistics of the 115 students in the five 8th grade classes. Table 1 shows that the average age was a little over 13, although age ranges from 12 to 15, mainly due to grade repetition. Three per cent of the students has repeated a grade at some point in his/her school career, 1.5% are new intake students that came from a different school and 57 students (49%) are female. The average primary school ability test score is 530 (test is scored on a scale between 500 and 550), with a minimum of 516 and a maximum of 545. Note that the primary school ability test score is missing for four students. This is explained by the fact that this test is not compulsory, although the large majority of primary schools use it. The average monthly income of the neighborhood in which the student is living (an indicator of socio-economic status—SES) is 2126 euros, but ranges from about 1500 to 3500 euros per month. This variable is taken from administrative records of the tax administration and reflects the average SES of all households within a particular neighborhood.

Statements on open data, ethics and conflict of interest

The data can be found at Research Gate, doi: 10.13140/2.1.4407.9049.

The data can be downloaded after sending a permission request to the owner of the data (authors of this paper), including a short project description.

Potential ethical issues are that control group students are denied access to potential beneficial educational innovations. However, given that these experiments had not proven their effectiveness, there also was a risk for treatment students. In essence, control students received the same education as the 8th grade students in the years before. Furthermore, at some points they even received more education than in the previous years, because they received the weekly formative tests in two of the three experiments. Also, because of (the registration of the results of) these formative tests, these students could be monitored better and the school could direct more attention to them in the year after the experiments, when necessary. Therefore, in the end there were no ethical issues that lead to specific action.

There is no conflict of interest for this work.

References

Banerjee, A. V., Cole, S., Dufflo, E. & Linden, L. (2007). Remedying education: evidence from two randomized experiments in India. *Quarterly Journal of Economics*, *122*, 3, 1235–1264.

Boekaerts, M. (2002). The online motivation questionnaire: A self-report instrument to assess students' context sensitivity. In P. R. Pintrich & M. L. Maehr (Eds), *New directions in measures and methods* Vol. 12 (pp. 77–120). *Advances in Motivation and Achievement*. New York: JAI / Elsevier Science.

Burns, M. K., Kanive, R. & DeGrande, M. (2012). Effect of a computer-delivered math fact intervention as a supplemental intervention for math in third and fourth grades. *Remedial and Special Education*, *33*, 3, 184–191.

Ghysels, J., Haelermans, C. & Prince, F. (2014). *The economics of information in human capital formation— evidence from two randomized experiments with low stakes tests in secondary education (TIER WP 14/27)*, Maastricht, TIER Working Paper Series.

Haelermans, C., Ghysels, J. & Prince, F. (2014). Increasing performance by differentiated teaching? Experimental evidence of the student benefits of digital differentiation. *British Journal of Educational Technology*. doi: 10.1111/bjet.12209.