Online Self-Assessment as a Learning Method

Daniel Gayo-Avello[1], Hortensia Fernández-Cuervo[2]

[1]Department of Informatics, University of Oviedo. Calvo Sotelo s/n 33007 Oviedo (SPAIN)
[2]Psychologist and Pedagogue
dani@lsi.uniovi.es, tensi_f@hotmail.com

Abstract. Algorithms and Programming Languages is a core subject in the BS Degree in Mathematics at the authors' university. Some of the students are very interested in computer science and computer programming but most of them find the subject quite hard. This situation is particularly stressed when concerning theoretic and, in fact, many students point at these contents as the main difficulty of the subject.

Because of this, the authors decided to explore new ways to improve the student learning of theoretical concepts. Thus, they analyzed the use of online self-assessment tools as a self-learning system. To perform this analysis two different kinds of tools were chosen and the authors developed an experiment to evaluate, on one hand, the possible use of self-assessment tools as self-learning systems and, on the other hand, to compare the tools to each other.

1. Introduction

Algorithms and Programming Languages (APL) is a first year core subject in the BS Degree in Mathematics with 12 credits (120 class hours). Half of these credits are devoted to theoretical concepts and the remainders are laboratory credits.

Some of the students in the subject have some experience in computer programming, but for most of them this subject is their first contact with computers so it turns out to be quite hard.

In the first classes the students learn main concepts (e.g., algorithm, processor, action, variable, etc). However, for many students the subject soon becomes harder (e.g., control structures, functions, recursion, etc).

Knowing the inherent difficulty of the subject for first year students, the teachers developed a website for the subject. In this website the students can get slides, lecture notes, problems and post questions to a newsgroup.

Three months after the beginning of the classes a poll was conducted to determine how the students felt about the subject, the teachers, their classmates and themselves. This poll implied three main conclusions:

1. The students felt very positive about both the subject and the teachers.
2. They thought the subject was important and difficult (specially theoretical concepts).
3. The students declared to study little or nothing, in special theoretic!

Such results made the teachers perplexed and pushed them to explore new ways to face the big problem in the subject: the learning of theoretical concepts.

2. Exposition of the Problem

While the teachers were conducting the poll, they wondered about offering some kind of self-assessment tool on the website based on multiple-choice questions. This way the students would have a tool that would allow them to assess the learning that they had acquired.

This approach was pretty simple but raised some questions. On one hand, it could convince the students, wrongly, that the subject would be assessed with multiple-choice questions whereas the exams would have a more practical nature. On the other hand, if such self-assessment questions were not changed with enough frequency they could lead to a simple learning of answers and not to a meaningful learning. Finally, if the tool only provided answers, but no explanations of them, it could dissuade the students from using the tool or even worst, disappoint them and deplete their motivation.

That's why different options from multiple-choice tools were looked for. After a pretty exhaustive research the authors found a self-assessment tool with a different approach: Duck[1].

Duck is a tool developed by the Biology Computer Resource Center at UMASS, Amherst. It's mainly a collection of PHP[2] scripts that allow us to provide "courses". A course is divided into one or more areas and each area includes some questions.

Three things distinguish Duck from multiple-choice tools: the kind of questions available, the assessment and the navigation style. Duck allows multiple-choice questions, short answer questions and extended response

[1] http://bcrc.bio.umass.edu/projects/duck/
[2] http://www.php.net

questions (these can be sent to the instructor for their evaluation). On the other hand, Duck does not provide any kind of score to the answers, since these are not wrong or right, but each answer has a feedback to clarify its meaning and solve the student doubts. Finally, Duck allows non-linear navigation, so the student can choose the most appealing questions, browse all the possible answers, go back to previous questions, etc.

Such characteristics made Duck a really interesting possibility to offer a self-assessment tool that would provide added value to the students.

Therefore, considering the need of new ways of improvement in the learning of theoretical concepts, the teachers saw the possibility of using self-assessment tools to reach such a goal, besides the need of analyzing both traditional multiple-choice tools and different tools such as Duck. This way, it would be possible to find if these tools could improve meaningful learning and, if this was the case, which one would be the most suitable.

3. Background

3.1. Assessment and Learning

There are many literature about the influence of assessment on learning; however, although most of the references state an improvement in the students learning, the authors have not found any reference which provided strong evidences to establish an unambiguous relationship between both phenomena.

For instance, reference [3] describes the use of a multiple-choice tool as an online assessment system; besides this, Davies states the positive results on the students. However, the main goal of this tool was not to improve the students learning but to assess this learning so the author does not assert anything about the relationship between assessment and learning.

Sly and Rennie talk about a similar experience: a computer based tool to perform formative assessment [6]. Again, no data about the influence of assessment on learning are provided.

In ITiCSE 2002 proceedings, Lapidot describes some experiences that show how not online self-assessment can act as a powerful motivation technique [4] but does not give any data about the influence of self-assessment on learning.

So, this paper intends to provide some evidences about the effectiveness of self-assessment as a learning method by means of a rigorous experiment to link unambiguously self-assessment and self-learning.

3.2. Operant Learning vs. Meaningful Learning

Two kinds of learning appear implicitly in self-assessment tools: operant learning [5] and meaningful learning [1].

Operant learning (or operant conditioning) is a learning process by which the behavior of an individual takes place based on its consequences (e.g., not to park in a fire lane to avoid a fine).

On the other hand, meaningful learning is acquired when the student establishes relationships between his previous knowledge and the new knowledge (e.g., when a teacher explains the way to calculate the triangle area beginning with the rectangle area).

If these concepts are translated to the experiment described in this paper, it can be found that traditional multiple-choice tools can fit into operant learning because each time the student answers a question the tool states whether the answer is right or wrong, inducing the reinforcement or extinction of such a behavior (the answer). On the other hand, Duck helps to reach a meaningful learning because when the student answers a question he does not get a score but a feedback about the question and the chosen answer; such feedback helps the student to elaborate his own reasoning and reach the most suitable solution.

4. Hypothesis Formulation

Thus, two questions should be considered: on one hand we had to found if online self-assessment tools could improve the students learning; on the other hand, we had to determined if a tool which provides feedback (meaningful learning) was better than multiple-choice tools (operant learning). Therefore, the experiment should try to prove the following hypothesis: “Use of web based self-assessment tools improves the students learning of computer programming theoretical concepts” and “The tool Duck is more suitable than multiple-choice tools”.

5. Research Methodology

5.1. Variable Selection

To perform the experiment only two variables were selected: the self-assessment tool would be the independent variable and the academic performance concerning only theoretical concepts would be the dependent variable. The main goal of the research would be to state the causality of the former in the later.

5.2. Population and Sample

The population chosen to perform the experiment comprised the students of APL in the authors' university for the academic year 2001/2002.

On this population a stratified random sampling was conducted; to do this the population was divided into subgroups considering the gender of the student and the number of years that he or she was retaking the subject.

Hence, the population comprised 23 women and 19 men. Besides this, only three people were retaking the subject by second year, the remainders were taking the subject by first year. This fact also divided the students by age, 18 years for first time students and 19 years for the repeat students.

Once the subgroups were set, a fixed number of individuals were randomly chosen from each subgroup and were assigned to the groups I, II and III. The nature of each of these three groups (experimental group or control group) was also randomly decided and turned out to be: I-control group, II-duck and III-traditional (the last ones were both experimental groups).

At the end of this process there was a sample of 24 individuals distributed in three groups, each of these groups had 4 women, 4 men and a repeat student.

Experimental mortality took place during the research in the three groups (although in no case it affected repeat students). Thus, the experiment finished with 6 individuals in the control group and 7 in both experimental groups.

5.3. Experiment Design and Procedure

The experiment is a bivalent design pretest-posttest with one control group –C– and one experimental group for each of the self-assessment tools (groups duck –D– and traditional –T–). Each group comprises 8 individuals randomly chosen. The three groups would take a pretest and a posttest. The control group would not receive treatment while the experimental groups would experiment two different levels of the independent variable: group D would use the tool Duck and group T would use a traditional multiple-choice tool.

The experimental sessions took place in one of the computer laboratories used to teach APL. These sessions were conducted in three Friday mornings in a row.

In the first session a Likert scale was administered to the students in order to find out their attitude to the subject. Later, they took a pretest to determine their academic performance in APL before the experiment (only theoretical concepts). Such a pretest included 20 items from the chapters "Functions and subroutines" and "Recursion", each item comprised 4 choices with only one right answer. The evaluation of this pretest penalized random answers by scoring 1/2 each right answer and –1/6 each wrong answer.

The students were notified about the way in which the pretest would be scored but they didn't know the real nature of the experiment. In fact, the students were persuaded to believe that they were taking part in a program to evaluate teaching quality so an "exam" environment did not bias their answers.

At this first session, just after the pretest, groups D and T received their first treatment session; the second one was administered the next Friday.

The treatment administered to group T consisted in the use, for 60 minutes in each session, of a multiple-choice tool. Such a tool comprised 40 items[3] with 4 choices each one. The tool allowed the student to know if he had provided a "right" or "wrong" answer besides calculating his final score or seeing all the right answers.

Group D used, for 60 minutes too, a tool based on Duck with the same items and choices. However, it was quite different from tool used by group T; on one hand, the answers were neither right nor wrong but their choice provided feedback about their suitability; on the other hand, the student didn't get any kind of score by using the tool.

At the end of this session both groups D and T were administered a user satisfaction test to find out their "feelings" about their respective tool.

To finish the experiment, a week after the conclusion of the treatment the three groups were administered a Thurstone scale to evaluate the students home work performance during the four weeks taken by the experiment. Besides this, the three groups took a posttest, which comprised the same items from the pretest although questions and choices appeared in a different order.

6. Discussion of Results

This experiment provided data to determine the students' attitude toward the subject, their theoretical knowledge before the treatment (pretest), their theoretical knowledge after the treatment or in its absence (posttest), and their performance in the subject.

From these data the authors had to conclude if there were significant differences between the mean scores reached by the three groups before and after the treatment. If these differences were statistically significant, and the three groups were equally capable and

[3] The items which appeared in both self-assessment tools, the pretest and the posttest were chosen from the chapters "Functions and Subroutines" and "Recursion". Only 25% from the items used by the tools appeared in the pretest and the posttest.

attitudinally equivalent then such differences only could be ascribed to the treatment. To perform the data analysis SPSS was used.

First of all, we had to found out if the three groups were attitudinally equivalent, equally capable (before the pretest) and their home work performance equivalent during the weeks taken by the experiment. To resolve this, the authors conducted a one-way ANOVA for each of the three measurements: attitude, capability and personal home work performance.

Later, we had to resolved if the performance shown by the control group in the posttest were similar to the one reached in the pretest, as well as comparing both experimental groups (D and T) with the control group and with pretest situation. The Student's T-test was conducted to perform all of these comparisons.

Finally, we had to check if group D (which had employed Duck and, presumably, acquired a meaningful learning) had obtained better results than group T (operant learning).

For the sake of brevity, numeric results are shown below in an appendix and here we will only provide a discussion of the conclusions drawn from them.

As it can be seen in the first part of the statistical analysis, the scores reached by the three groups in all the measurements (pretest, posttest, attitude and home work performance) follow a normal distribution and show homogeneous variances. This way, the requirements to apply both the Student's T-test and ANOVA are fulfilled.

Applying ANOVA to the results obtained in the pretest, the posttest, and the attitude and home work performance tests showed that the three groups were statistically equivalent; that is, their attitude and capability were pretty similar, and during all the time taken by the experiment the students from the three groups studied mostly the same (that is, little).

What this means is that the noticeable differences in the posttest cannot be ascribed to capability or attitudinal differences, nor a different effort from the students, they can be only credited to the only variable that allows us to distinguish the groups: the self-assessment tool used in the experiment.

The Student's T-test proved that in the posttest both groups D and T reached a significant improvement with regard to group C (control) and to their respective previous situation in the pretest; on the contrary, group C got worse slightly although not significantly (that is, it remained with no changes).

Therefore, the first hypothesis is true: the use of web based self-assessment tools improves the students learning of computer programming theoretical concepts.

With regards to the second hypothesis, the advantage of Duck over traditional multiple-choice tools, couldn't be proven because both experimental groups (D and T) didn't reach significant differences between them. However, this doesn't mean that both tools are equally suitable; it is necessary to perform more experiments to finally prove or reject such hypothesis.

The graphics included in the appendix show in a highly intuitive way the experiment results.

7. Conclusions and Future Work

With such results it can be stated that the use of Web based self-assessment tools help the students to learn theoretical concepts about computer programming.

The question about the best suitability of multiple-choice tools or feedback providing tools remains open.

It is true that the satisfaction rate was higher in the group that used a feedback providing tool, such as Duck, that in the group which used a multiple-choice tool; however, the results reached in the posttest by both groups where statistically equivalent. Despite of this, it must be said that the students would employ such tools in a voluntary basis. Because of this, it is likely that a real application would show quantifiable differences in the results reached by the students depending on the chosen tool. On the other hand, it must be considered that, in general, only the best students tend to benefit the most [2] from initiatives like the one described in this paper.

For such reasons, once the suitability of self-assessment tools as a learning method has been stated, the authors want to perform a new research in order to find out which of these two tools (multiple-choice of feedback) is the most suitable under real conditions.

References

[1] Ausubel, D.P., Educational Psychology. A cognitive View, Holt, Rinehart & Winston, 1968.

[2] Carver, C.A., and Howard, R., An Assessment of Networked Multimedia and Hypermedia, in Proceedings of the 1995 IEEE/ASEE Frontiers in Education Conference, Nov. 1995, 2c5.17-2c5.21.

[3] Davies, P., Learning through assessment OLAL…. On-line Assessment and Learning, in Proceedings of the 3rd Annual CAA Conference, June 1999, Loughborough University, pp. 75-88.

[4] Lapidot, T., Self-Assessment as a Powerful Learning Experience, in Proceedings of ITiCSE'2002, Innovation and Technology in Computer Science Education, June 2002, ACM Press, pp. 198-198.

[5] Skinner, B.F., The Behavior of Organisms: An Experimental Analysis, B.F. Skinner Foundation, 1938 originally, reprinted 1991 and 1999.

[6] Sly, L., and Rennie, L.J., Computer managed learning: Its use in formative as well as summative assessment, in Proceedings of the 3rd Annual CAA Conference, June 1999, Loughborough University.

Appendix: Statistical Analysis

We are using 0.05 significance level

Pretest scores: Shapiro-Wilk normality test

Group	W	Sig.	Conclusion
Control	0,938	0,559	0,938 ≅ 1 y 0,559 >> 0,05 → Normal distribution
Duck	0,898	0,365	0,898 ≅ 1 y 0,365 >> 0,05 → Normal distribution
Traditional	0,947	0,672	0,947 ≅ 1 y 0,672 >> 0,05 → Normal distribution

Pretest scores: Levene's test for equal variances

Levene Statistic	df1	df2	Sig.	Conclusion
0,879	2	19	0,431	0,431 >> 0,05 → Variances are equal

Pretest scores equivalence: one-way ANOVA

	Sum of Squares	Degrees of freedom	Mean Square	F	Sig.	Conclusion
Between groups	17,537	2	8,769	2,645	0,097	0,097 > 0,05 → Equivalent capabilities
Within groups	62,977	19	3,315			
Total	80,514	21				

Attitudinal scores (Likert scale): Shapiro-Wilk normality test

Group	W	Sig.	Conclusion
Control	0,972	0,899	0,972 ≅ 1 y 0,899 >> 0,05 → Normal distribution
Duck	0,967	0,857	0,967 ≅ 1 y 0,857 >> 0,05 → Normal distribution
Traditional	0,921	0,471	0,921 ≅ 1 y 0,471 >> 0,05 → Normal distribution

Attitudinal scores (Likert scale): Levene's test for equal variances

Levene Statistic	df1	df2	Sig.	Conclusion
0,391	2	19	0,682	0,682 >> 0,05 → Variances are equal

Attitudinal scores (Likert scale) one-way ANOVA

	Sum of Squares	Degrees of freedom	Mean Square	F	Sig.	Conclusion
Between groups	0,305	2	0,153	0,014	0,986	0,986 > 0,05 → Equivalent attitudes
Within groups	200,286	19	10,541			
Total	200,591	21				

Home work performance scores (Thurstone scale): Shapiro-Wilk normality test

Group	W	Sig.	Conclusion
Control	0,930	0,522	0,930 ≅ 1 y 0,522 >> 0,05 → Normal distribution
Duck	0,864	0,214	0,864 ≅ 1 y 0,214 >> 0,05 → Normal distribution
Traditional	0,909	0,416	0,909 ≅ 1 y 0,416 >> 0,05 → Normal distribution

Home work performance scores (Thurstone scale): Levene's test for equal variances

Levene Statistic	df1	df2	Sig.	Conclusion
1,345	2	17	0,287	0,287 >> 0,05 → Variances are equal

Home work performance scores (Thurstone scale): one-way ANOVA

	Sum of Squares	Degrees of freedom	Mean Square	F	Sig.	Conclusion
Between groups	8,568	2	4,284	1,013	0,384	0,384 > 0,05 → Equivalent home work performance
Within groups	71,879	17	4,228			
Total	80,448	19				

Posttest scores: Shapiro-Wilk normality test

Group	W	Sig.	Conclusion
Control	0,912	0,441	0,912 ≅ 1 y 0,441 >> 0,05 → Normal distribution
Duck	0,908	0,410	0,908 ≅ 1 y 0,410 >> 0,05 → Normal distribution
Traditional	0,985	0,978	0,985 ≅ 1 y 0,978 >> 0,05 → Normal distribution

Posttest scores: Levene's test for equal variances

Levene Statistic	df1	df2	Sig.	Conclusion
0,341	2	17	0,716	0,716 >> 0,05 → Variances are equal

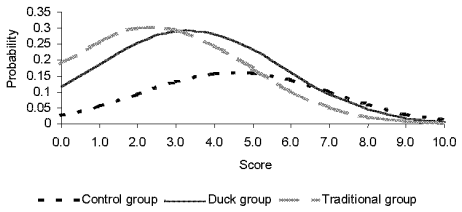
Posttest scores improvement: Independent Samples Student's T Test

Groups	T	Degrees of freedom	1-tailed Sig.	Conclusion
Duck vs. Control	3,027	11	0,0060	0,0060 << 0,05 → Significant improvement
Traditional vs. Control	2,965	11	0,0065	0,0065 << 0,05 → Significant improvement
Duck vs. Traditional	0,265	12	0,796	0,7960 >> 0,05 → **NON** Significant improvement

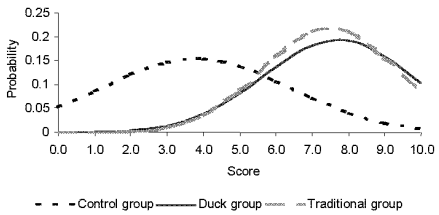
Posttest scores improvement: Paired Samples Student's t Test

Groups	T	Degrees of freedom	2-tailed Sig.	Conclusion
Duck vs. Duck	11,664	6	0,000	0,000 << 0,05 → Significant improvement
Traditional vs. Traditional	20,457	6	0,000	0,000 << 0,05 → Significant improvement
Control vs. Control	-0,706	5	0,512	0,512 >> 0,05 → **NON** Significant improvement

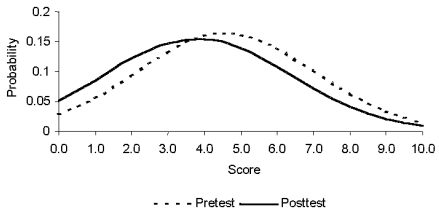
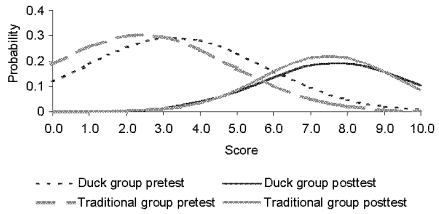
Pretest score distribution



Posttest score distribution



Control group evolution



Experimental groups evolution