Student Self-assessment in a Programming Course Using Bloom’s Revised Taxonomy

Satu Alaoutinen
Lappeenranta University of Technology
P.O. Box 20
53851 Lappeenranta, Finland
satu.alaoutinen@lut.fi

Kari Smolander
Lappeenranta University of Technology
P.O. Box 20
53851 Lappeenranta, Finland
kari.smolander@lut.fi

ABSTRACT
This study presents a simple student self-assessment tool that can be used to motivate learning and to follow a student’s progress. The tool is a survey questionnaire that uses Bloom’s Revised Taxonomy as the base for its scale. The results show that students can place their knowledge along the taxonomy-based scale quite well, and they feel that it could help their learning. It also provides the teacher with a more objective basis for measuring the level of knowledge gained than the general scales. The scale is designed for a programming course, but with modifications it can be used in other classes too.

Categories and Subject Descriptors
K.3.2 [Computers and Education]: Computer and Information Science Education—Self-assessment

General Terms
Experimentation

Keywords
Self-assessment, Bloom’s Revised Taxonomy, Programming

1. INTRODUCTION
In our university students are being pushed to finish their studies faster, and programming, along with mathematics, has been a topic that has caused problems for many students. In many cases, basic programming is the last course that keeps them from graduating, while it was one of the first courses taught in our university. We have been seeking ways to improve our teaching of programming. The language taught in our fundamental programming course has already been changed, and the technical infrastructure improved, so now it is time to concentrate on teaching aspects [8].

One key issue in teaching is motivating students. By giving the students goals to achieve and a means to follow their progress, learning can be meaningful, especially in compulsory fundamental courses which students are required to pass. Recently, many universities have started to require learning outcomes [9] to their courses. Learning outcomes are based on the level of knowledge of the student and they should define clearly what is expected from a student. Therefore, they can be used to set goals for students. Learning outcomes also provide a basis for course assessment [1].

One form of assessment is self-assessment. Studies show that when students evaluate their own work or the work of others their results improve [10]. There is a problem, however, in assessing how well students’ evaluations match their actual skills and knowledge.

If students could evaluate their own learning using a simple survey instead of a more complicated tool, more teachers would use it. Students could use self-assessment as a means of following their own development, and when combined with learning outcomes, to set clear goals for the course. We were interested to know whether students are able to evaluate their own learning and to recognize the corresponding level of knowledge from Bloom’s Taxonomy. If they can do that, it would be easy for a teacher to follow the learning process by asking students to fill in a simple self-assessment survey several times during the course. When combined with course learning outcomes, the survey could provide students with the means to follow their own development.

2. BLOOM’S REVISED TAXONOMY
Many studies use Bloom’s Taxonomy [3] (referred to later as the original Taxonomy) or its revised version [1] (referred to later as the revised Taxonomy) as their reference point or framework. This taxonomy describes knowledge hierarchically and cumulatively. Attaining the higher levels requires having mastered the levels below. The original Taxonomy defines three domains of knowledge (cognitive, affective, and psychomotor) and it separates domains into six levels. Since there was a debate about whether or not the order of the three highest levels was correct, a new revised version of the Taxonomy was created by Anderson et al. They combined the cognitive process dimension with the knowledge dimension, which consists of factual, conceptual, procedural, and metacognitive knowledge, as depicted in Figure 1. Quite often the knowledge dimension is left unused and the cognitive process dimension is used in its place.

The simplest form of knowledge is rote learning, but it is a foundation of deeper knowledge. There are always some
A programming project requires evaluation skills and finding skills. For example, choosing a sorting algorithm in a level than applying and it requires analyzing and evaluating before they can be applied. Creating is at a higher level than applying and it requires analyzing and evaluation skills. For example, choosing a sorting algorithm in a programming project requires evaluation skills and finding reasons for program malfunctions requires analysis skills.

3. RELATED RESEARCH

Assessment is a vital part of an educational system, and it has been much studied. Self-assessment is a good way to improve learning results and it also provides motivation for learning [10]. Despite its effectiveness, self evaluation is not widely used as an instructional tool. The reason for this could be the extra work required to create a reliable self-assessment system. For example, Sung et al. [12] have developed a web-based self- and peer-assessment system. According to their study, the students are quite skilful in assessing both their own and others’ work. They report that the assessment skills improve through practice, and that the students were able to improve their work via assessment.

Some studies try to verify the reliability of self-assessment. Crews and Turner [5] compare the students’ knowledge of telecommunications against their own assessment. They test the actual knowledge using 100 randomly selected multiple-choice questions from a course textbook sorted into 5 telecommunications dimensions: 1) networks and infrastructure, 2) hardware and software, 3) management and skills, 4) performance and security, and 5) standards and protocols. The students evaluated their level of knowledge in each of the dimensions using a five-step Likert scale, which ranges from very low (1) to very high (5). The small sample size prevents any generalization, but they notice a common trend that the students know more than they imagine they do.

The creators of the taxonomies, both the original and revised one, give examples of how to apply it. The examples cover many disciplines, but unfortunately computer science is missing from the list. However, some teachers have created proposals for how to use them in computer science; for example, in programming. Thompson et al. [13] have created guidelines for how to apply the revised taxonomy to programming assessment. They provide hints as to what kinds of questions are needed at each level, and provide some concrete examples as well. Hernan-Losada et al. [6] proposes suitable programming tasks for the levels ranging from comprehension to synthesis in the original taxonomy.

Another study by Buckley and Exton [4] maps software maintenance tasks using the original taxonomy, and demonstrates how knowledge of software may differ at various levels of the cognitive process dimension. Xu and Rajlich [14] study the cognitive process during program debugging. They found that searching for the reason for an error requires using all levels of cognitive knowledge in the Bloom’s Taxonomy. Debugging has always been difficult for beginning programmers, and this explains why. They suspect that novices have to spend their time just absorbing knowledge and understanding it.

Hwang et al. [7] have used the Bloom’s taxonomy as a reference point to make sure that all levels of thinking are included in programming learning activities designed for a new web-based programming learning environment which also includes the possibility of peer-assessment.

There are also other taxonomies used for assessment purposes which are more or less derivative of Bloom’s taxonomy, but they are not widely used. The SOLO-taxonomy [2] has been used as an alternative computer science assessment, but it concentrates on the content of the responses instead of the cognitive performance. It is also difficult to apply it to self-assessment because it requires knowledge about the correctness of the student’s own answers.

In summary, there are many tools for student self-assessment, but they are rather complex for our purposes. Some studies have used simple methods, like Likert scales, for self-assessment. They are simple enough, but lack objectivity, because the same things can have different meanings for different people (i.e. I manage recursion well). Students’ actual knowledge has been tested in different ways. A laborious way is to use special test questions, while an easier way is to use exam questions. For our purposes, it is best to use exam results as the verification method. In order to determine the level of students’ knowledge, Bloom’s Revised Taxonomy seems to be an applicable method because it makes the distinction between different levels of knowledge. In addition, there are precedents available for how to use it to study programming-related processes.

4. RESEARCH METHOD

This is a survey study, which uses a web-based questionnaire. The study was conducted in one introductory C-programming course in which 87 students were actively involved. The students were from different departments and included a mixture of freshmen and advanced students. In our system, students can leave some courses open and finish them in later years.

As was noted earlier, the creators of the revised taxonomy did not provide any examples for how to use it within the computer science discipline. However, we were able to find two articles in which the researchers have addressed this
shortcoming. Hernan-Losada et al. [6] and Thompson et al. [13] have presented sample problems and questions for programming to be used at different levels of the taxonomy. For example, for level three they suggest tasks such as “applying a known formula or method in a new situation”, or “that the process and algorithm or design pattern is known to the learner, and both are applied to an unfamiliar problem”. The assessment scale used in the survey generalizes those questions and problems. So, the level-three task in our scale is: “I can list cases, when that command/concept can be used. I can apply an example to a different problem”.

Using a web questionnaire, the students were asked to give their student ID and rank their skills in different programming topics by answering the following question:

Evaluate for every item at what level (1..6) your knowledge best fits. The scale 1..6 means:

1. I can list related commands/concepts.
2. I can explain what the command/concept means.
3. I can apply an example to a similar problem.
4. I can list cases when the command/concept can be used. I can apply an example to a different problem.
5. I can explain the meaning of the command/concept in its context, why it is there.
6. I can ensure the correct use of the command/concept.

Tick the appropriate box for each item.

The whole list of course topics is visible in the results section (Figure 2). The list actually has 22 items for C-programming language and programming in practice, covering data types and operators as well as dynamic memory allocation and version control.

Because the assessment scale is ordinal, but the distance between the steps is not known, we calculated a median for each of the students based on his/her answers to the 22 questions that could be compared to his/her exam grade. We used exam grades as a verification point simply because they were easily available and gave the results we needed. From the course enrolment data, the student’s year in school, department, and gender were readily available.

The students were also asked three open-ended questions:

1. How was it to answer this questionnaire?
2. If this questionnaire was published in the beginning of a course combined with the learning outcomes, would it help learning?
3. Can you think other ways to use the questionnaire?

5. RESULTS

50 students answered the survey and 48 answers were valid. Two answers were missing the student ID, so they could not be combined with their exam result. In addition, four more answers were removed. Two were considered as outliers (box plot), one was a row of ones with the comment that there was not sufficient time to concentrate on the questions, and one student did not make any effort in the exam.

So, this left 44 answers to be used, representing about 51 % of the active group.

Figure 2 shows how the students have answered the questions for different items. The y-axis presents the level of the taxonomy ranging from remembering (1) to creating (6); the x-axis lists the topics handled in the course. There are two set of marker lines, a median based on what the students ticked for each item, and the lowest assessment scores. The lowest scores are not all from the same student. The topics along the x-axis are arranged in the same order as they were taught. As can be seen, the fundamental things, like data types, operators, input, and output, are well mastered by most of the students. The median is at the highest level, creating. When moving to the right along the x-axis the level of mastery gets lower. Things get more difficult and there is not as much time to practice. The most difficult things in C-language, pointers and dynamic memory handling, stick out at their own level. Most of the students think that they can use pointers and allocate memory if they have supportive examples available. This corresponds quite well to the experiences the teachers have from the class exercises. Figure 3 is a scatter plot of each student’s survey median and his/her exam result. The dots represent the students. Each student has a survey result/exam result pair,
and many students can have the same results. The relative size of the diamond tells how many students just received that particular grade-median combination. For example, of the students who received a 4 on the exam, 2 of them received a 4 as the median score on the self-assessment, 9 students received a 5 on the self-assessment, and 3 students received a 6. The Pearson product moment correlation between the survey median and the exam grade is 0.5429 and the p-value of that correlation is 0.0001. The figure shows a rising trend in self-assessment as the grades improve. When looking at the results more closely, the older students seem to be better at assessing themselves (figure 4). In our test course, the threshold was between the second and third year. For first and second year students the correlation was 0.58 with a p-value 0.0039 (N=26), and for the older students it was 0.84 and 0.0000 (N=18), respectively.

Figure 4: Scatter plot - 1st and 2nd year (left)- 3rd year and beyond (right).

There were also differences in the scores between the different majors. The largest group of students came from computer science, electrical engineering, and industrial management. The industrial management students were the most accurate in self-assessment (r=0.66), while the worst were the computer science students (r=0.43). Only 4 of the respondents were female, so gender could not be used in the analysis, even though their assessments were quite accurate. None of these correlations were statistically significant.

The three open-ended questions were analyzed by simple open coding [11]. The first question dealt with the answer process itself. 32 out of 44 students (73%) answered. 21 of them considered it easy to answer. 13 did not give any opinion. 2 mentioned that the scale was clear, but 6 considered the scale peculiar, too general, difficult, or confused. Our students used to use ordinary Likert-type scales in assessing courses or their own learning. The taxonomy-based scale is new to them; as suggested by one comment: “Slightly extraordinary choices”. There are other answers too, which indicate that the difficulty stems from the novelty of the scale. In addition, one student commented, rather bluntly, that “it is difficult to evaluate one’s own knowledge”. Based on the comments, it looks like the scale is understandable for many of the students, and that a lack of experience caused many of the difficulties.

Table 1 shows student responses to Questions 2 and 3. The Question 2 described the intended use of the questionnaire and therefore guided the responses. Two thirds of the answers were positive or neutral and one third did not think that the questionnaire would help learning. One even stated that relative progress in learning cannot be evaluated without a formal exam. Some of the negative answers indicate that the question itself was badly formulated. For example, immediately after making the comment that the questionnaire would not help learning, the same student described an alternative way of using the questionnaire, which was essentially our idea put in other words: “the assessment scale could be used along with the course as a table for measuring the progress of own skills”. Therefore we do not consider this question and the answers to it to be fully reliable. The second largest group of practical suggestions in the Question 3 was that it would help to see what the course contains. This does not fully make sense, because the same list of topics is already available on the course web site. The largest group of responses agreed that it would help in goal setting if the required knowledge levels are known from the beginning of the course.

There was no connection between the answers. A student might feel that the questionnaire was difficult to answer, but also saw that it could help with learning, or vice versa. Neither did the exam results connect readily to any specific answers.

6. DISCUSSION

Traditional scales used in self-assessment, like in [5], are subjective and those studies indicate that such self-assessment surveys are not reliable. Replacing Likert-type scales with a taxonomy-based scale makes assessment more objective by defining the requirements for each level. We wanted to use a taxonomy scale for assessment of levels of knowledge, but only two taxonomies, Bloom and SOLO, have been widely used for assessment purposes and only Bloom’s Taxonomy concentrates on the depth of knowledge. Because of this, we decided to use Bloom’s Revised Taxonomy as a basis for our self-assessment scale.

This study shows that students in general seem to be quite accurate in assessing their own knowledge. There is a statistically significant correlation between the self-assessment results and the course exam results. In particular, the lower levels of the taxonomy seem to be easy to distinguish. Even though the correlation exists, it has to be noted that there are plenty of students, especially younger students, who over- or underestimate their skills. For them, this scale does not provide a realistic image of their knowledge. One reason is probably the lack of experience. We gave this survey to our students without any prior preparation or explanation. The older students have more experience in assessing and reflecting upon their own learning, so they could grasp this

<table>
<thead>
<tr>
<th>Answer category</th>
<th>Number of answers</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>No answer</td>
<td>11</td>
<td>24</td>
</tr>
<tr>
<td>No help at all</td>
<td>14</td>
<td>31</td>
</tr>
<tr>
<td>Helps seeing what exactly the course contains</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>Helps in preparing for the exam</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>Tells the teacher, what has been difficult</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Start-end comparison</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>Goal setting</td>
<td>7</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 1: Question 2 answer categories
new kind of assessment scale better. But, as pointed out by Sung et al. [12], assessment skills improve through practice. If the scale had been explained before its first usage, the results will most likely have been more accurate. Some differences were also found between different departments but they were not so significant.

The topic-based grouping of answers (Figure 2) seems to provide a realistic image of students’ learning in general. The easiest topics as well as the most difficult ones are clearly visible and correspond to our view. Therefore, it could be a useful tool when a teacher wants to follow the students’ progress. A teacher can, for example, ask the students to mark their knowledge levels every three weeks and follow how the knowledge level rises after practice. If a particular topic remains at a low level of knowledge, even when it should not, then something needs to be done. In addition, when learning outcomes have been defined, the right target level is easy to define.

Some of the students felt that the scale was not clear. That is not surprising, because some sentences at higher levels of the scale were difficult to formulate. The students’ suggestions for how the tool could be utilized were partly expected and partly confusing. Their suggestions covered goal setting both for a course and at an individual level, start-end comparison, and class advancement monitoring. These answers tell us that we are on the right track. Yet, it would have been interesting to know what some of the students meant by suggesting “Helps seeing what exactly the course contains”. It seems that they have not visited the course web site at all. This would have needed a follow-up question.

The ability to draw generalizations based on this study can be improved by extending the sample size and by including several courses and universities in the sample. In this study, the sample size was relatively small and so far this scale has only been tested in one programming course.

7. CONCLUSIONS

In this study, we created a general assessment scale for student self-assessment in a programming course. The levels of that scale were taken from the Bloom’s Revised Taxonomy by making generalizations based on the descriptions of the knowledge levels in the taxonomy and applying them to programming concepts. The results show that the scale used in this form gives quite a good general picture of students’ knowledge level.

The topic based categorization of answers is a useful tool for a teacher. It tells clearly what topics are difficult and require further attention. In addition, it also tells precisely what level of knowledge most of the students have achieved. This represents an improvement over the assessment scales ranging, for example, from poor to excellent, which are fully subjective. After rewording, and with proper guidance, the scale could be a useful tool for the students also. The students could set goals for themselves and follow their development or use it as an aid while preparing for an exam to see what is required at each level. The over- and underestimations of knowledge are causing some trouble, but practicing assessments helps in that matter.

For the teacher, the assessment will work as it is, but requires learning outcomes alongside it to be useful for the students. Some teachers already use learning outcomes in their courses, and for the rest they should also be recommended. This tool is a motivation for using learning outcome descriptions as well.

8. REFERENCES


