Extending the Engagement Taxonomy: Software Visualization and Collaborative Learning

NIKO MYLLER, ROMAN BEDNARIK, and ERKKI SUTINEN
University of Joensuu
and
MORDECHAI BEN-ARI
Weizmann Institute of Science

As collaborative learning in general, and pair programming in particular, has become widely adopted in computer science education, so has the use of pedagogical visualization tools for facilitating collaboration. However, there is little theory on collaborative learning with visualization, and few studies on their effect on each other. We build on the concept of the *engagement taxonomy* and extend it to classify finer variations in the engagement that result from the use of a visualization tool. We analyze the applicability of the taxonomy to the description of the differences in the collaboration process when visualization is used. Our hypothesis is that increasing the level of engagement between learners and the visualization tool results in a higher positive impact of the visualization on the collaboration process. This article describes an empirical investigation designed to test the hypothesis. The results provide support for our extended engagement taxonomy and hypothesis by showing that the collaborative activities of the students and the engagement levels are correlated.

Categories and Subject Descriptors: K.3.2 [**Computer Science Education**]: Computer & Information Science Education—*Computer Science Education*

General Terms: Algorithms, Experimentation, Human Factors

Additional Key Words and Phrases: Program visualization, collaborative learning, engagement taxonomy

The research of N. Myller and R. Bednarik has been funded by the East Finland Graduate School in Computer Science and Engineering (ECSE). A part of this paper was written while N. Myller was a visiting lecturer at the Tumaini University, Iringa University College, Iringa, Tanzania under the 5ARTS+ project funded by CIMO.

Authors' addresses: N. Myller, R. Bednarik, and E. Sutinen, Department of Computer Science and Statistics, University of Joensuu, P.O. Box 111, FI-80101 Joensuu, Finland; email: nmyller@cs.joensuu.fi; M. Ben-Ari, Department of Science Teaching, Weizmann Institute of Science, Rehovot 76100, Israel.

ACM Reference Format:
Myller, N., Bednarik, R., Sutinen, E., and Ben-Ari, M. 2009. Extending the engagement taxonomy: Software visualization and collaborative learning. ACM Trans. Comput. Educ. 9, 1, Article 7 (March 2009), 27 pages. DOI = 10.1145/1513593.1513600.
http://doi.acm.org/10.1145/1513593.1513600.

1. INTRODUCTION

When *algorithm visualization (AV)* and *program visualization (PV)*[1] were first introduced more than two decades ago, they seemed to be a silver bullet that could solve difficult problems related to the teaching and learning of programming, data structures and algorithms. However, the mixed results of empirical evaluations have made the benefits of visualization tools as teaching and learning aids questionable [Hundhausen et al. 2002]. Therefore, researchers have begun to seek explanations for the mixed results in order to discover the conditions under which visualization tools can actually achieve improvements in learning.

In a meta-analysis of the research on AV, Hundhausen et al. [2002] concluded that the activities performed by students and their engagement seem to be more important than the subject content or the graphic elements of the visualization. The findings led to the analysis of different engagement levels between the user and the visualization tool, resulting in the *engagement taxonomy (ET)* described by Naps et al. [2002]. The main assumption of the taxonomy is that the level of engagement between the user and the visualization affects the learning of the individual student. Since its introduction, the ET has guided the research and development of SV tools, and several studies have utilized the framework [Grissom et al. 2003; Naps and Grissom 2002].

Collaborative learning and pair programming have become accepted and popular methods in computer science education [Hundhausen and Brown 2008; McDowell et al. 2006; Simon et al. 2004; Nagappan et al. 2003; Hundhausen 2002; Williams et al. 2000]; as the use of visualization tools increases [Naps et al. 2002], they appear more and more often in situations of collaborative learning. This combination introduces new challenges and possibilities that are different from the ones related to individual learning with visualization. On the one hand, successful collaboration requires the communication of knowledge and new ideas between group members, as well as the coordination of joint work, but it is not clear how visualization affects these issues [Suthers and Hundhausen 2003]. On the other hand, visualization tools themselves can create a context for collaboration by providing a shared external representation that can initiate negotiations of meanings; they can also become a reference point for explaining ideas or resolving misunderstandings. The mutual influence of visualization and collaboration on each other is likely to be relevant for their joint analysis through means such as the engagement

[1] We will use *software visualization (SV)* to refer to both of these subfields.

taxonomy; therefore, it is unlikely that the ET for collaborative learning will be same as it is for individual learning.

Visualization tools that are used during collaboration can be divided into two categories: *information visualization* such as concept maps or SV tools that visualize programs or data structures, and *augmenting visualization* such as social and group awareness software. Augmenting visualization tools are known to enhance the process and the outcomes of the collaboration [Janssen et al. 2007]. In this article, we concentrate on information visualization, particularly on SV, because the effects of its content and form on collaboration have not been investigated in depth. There are few theories of collaborative learning that apply to information visualizations and few tools that support the collaborative learning [Suthers and Hundhausen 2003; Suthers et al. 2003]. Thus, users have little guidance from the research literature as to how to adjust the use of a tool meant for an individual to the different needs of collaborative settings [Bryant et al. 2005].

In this paper, we first review literature related to the collaborative use of visualization tools (Section 2). We then extend the engagement taxonomy with 1) new levels to an *extended engagement taxonomy (EET)* that identifies finer levels of distinctions in the engagement in both individual and collaborative learning (Section 3.1) and 2) a hypothesis that takes into account aspects of collaborative learning process (Section 3.2).

Our hypothesis is that the higher the level of engagement between learners and the visualization tool, the higher is the positive impact of the visualization on the collaboration process. To test this hypothesis, we present an empirical study, in which we investigated the activities of groups of students at different engagement levels as supported by two visualization tools (Section 4). In the study, we analyzed the interactions between students, and between the students and a visualization. Section 5 presents the results and provides evidence for our hypothesis. In the final section, we discuss the implications of our findings on the future research and development of collaborative learning with software visualization.

2. RELATED WORK

This section is an integrated survey of the relevant previous work on visualization, collaborative learning and the engagement taxonomy.

2.1 Software Visualization and the Engagement Taxonomy

In an attempt to describe the mixed results of previous research on SV in education, the engagement taxonomy was introduced by Naps et al. [2002]. Its purpose is to describe the different forms of engagement that a visualization tool can promote, and to provide testable hypotheses about the use of visualizations in the teaching and learning of computer science. The central idea of the taxonomy is that higher-level engagement between learner and the visualization results in better learning outcomes. The ET consists six levels of engagement between the user and the visualization (see Table I).

Table I. The Engagement Taxonomy

No viewing	There is no visualization to be viewed.
Viewing	The visualization is only looked at without any other form of engagement.
Responding	Learners are presented with questions related to the visualization.
Changing	Modification of the visualization is allowed, for example, by varying the input data set.
Constructing	Learners are expected to create their own visualization of a program or an algorithm.
Presenting	Learners present visualizations to others for feedback and discussion.

When there is no visualization to look at, the engagement is, of course, at its lowest level. Passive viewing of a visualization seems to improve learning outcomes very little, if at all, even when compared with the no viewing level [Hundhausen et al. 2002; Naps et al. 2002; Naps 2005]. One can conclude that there should be an active component in the learning process in order to enhance learning with visualization. That is, the viewing of a visualization should be combined with activities at the higher levels of engagement: responding, changing, constructing or presenting. To our knowledge, there are few empirical comparisons between these forms of active engagement on learning outcomes [Naps et al. 2002; Naps 2005]. In light of current research, the taxonomy forms a three-level hierarchy: no engagement, passive engagement, and active engagement [Naps 2005].

The ET has been used in the development of AV tools, and studies have validated its applicability [Grissom et al. 2003; Naps and Grissom 2002; Myller et al. 2007b].

Other studies—although not using the ET—have shown that visualizations enhance learning; for example, Ben-Bassat Levy et al. [2003] found that students who actively used the Jeliot program animation system improved their learning results compared with a control group that did not use Jeliot. Also relevant is the research in educational psychology and multimedia learning, which has found a positive effect of the interactivity in multimedia on learning outcomes [Evans and Gibbons 2007]. These studies have concentrated on changes in the learning outcomes when visualization is used. In this article, we extend the scope of such research in order to investigate the impact of the engagement with visualization tools on the learning process.

The effects of visualization on the learning process have been also researched, although not in a collaborative environment. For example, Ebel and Ben-Ari [2006] showed that program visualization increases the attention of students to the material being taught. We believe that this could also hold in a collaborative environment: as students' attention increases, they will be able to concentrate better on the collaborative activities, but it should be tested in the future experiments.

2.2 Use of Visualizations in Collaborative Learning

The work of Roschelle [1996] is considered seminal for the whole field of *computer-supported collaborative learning (CSCL)*. Roschelle developed the *envisioning machine*, a software tool for studying mechanics that students use to manipulate simple diagrams related to velocity and acceleration. He investigated how pairs of students used the tool, and he analyzed the learning outcomes as well as the processes that led to those outcomes. He recognized that learning tools for collaboration should be designed to support communication, rather than merely to present the underlying model as accurately as possible. Roschelle [1996] gives a number of guidelines in order to achieve this goal; the final one is "one should design activities which actively engage students in doing and encounter [sic] meaningful experiential feedback as a consequence of their actions" (p. 14). The analysis of the interaction between the external presentation and users was also identified as a key research area by Scaife and Rogers [1996]. The work of Roschelle [1996] and Scaife and Rogers [1996] reflects the idea that the engagement with visualizations affects collaborative learning, and we build on this in the article.

Suthers and Hundhausen [2003] compared the effects of different representations (matrices, graphs, text) when students collect and analyze data, form hypotheses and investigate their evidential relations, both in a face-to-face and in a distance context [Suthers et al. 2003]. They found that there were differences in the guidance that different representations give to the collaboration, especially to discussions, and that the different learning situations (face-to-face or distance) affect the usage of the representations [Suthers et al. 2003]. However, they did not find differences in the performance of the students, but only in the way the students used and discussed the representations. It could be argued that the differences in the study process should have an effect on students' learning only in a long run, and therefore were not detectable in the laboratory setting.

The research described in this section shows that visualization and the kinds of interactions it drives have an effect on the collaboration process and could affect the collaboration outcomes.

2.3 Research on Software Visualization in Collaborative Learning

Although a plethora of software visualization tools have been developed and empirical studies carried out, there have been only a few tools and studies relating to the collaborative use of SV.

Myller et al. [2007b] studied learning outcomes after students had collaboratively learned about the concept of binary heap with the help of either animation (ET level: viewing) or algorithm simulation (ET level: changing). Although they did not find statistically significant differences in the performance between the groups, the groups that used algorithm simulations performed consistently better in a post-test, compared with the groups that viewed just the animations. In a replication of the study, a statistically significant difference was found in favor of the algorithm simulation group, a finding that supports the applicability of ET in the context of collaborative learning with visualization [Laakso et al. 2008].

Hundhausen [2002] studied the collaborative aspects of AV construction and presentation, and concluded—as did Roschelle [1996]—that the fidelity of the visualization can be compromised in favor of meaningful interactions between students. This led into the development of ALVIS, a visualization tool that supports construction and presentation of AVs [Hundhausen and Brown 2007]. In an experiment, they compared ALVIS as a tool to writing programs for algorithms to a text editor, and then the use of ALVIS for visualization construction and presentation to simple art supplies [Hundhausen and Brown 2008, 2005]. Students worked in pairs and were asked to write an algorithm in the SALSA language supported by ALVIS, construct a visualization of that algorithm, and present it to the instructor and the other students. It was found that pairs of students who used ALVIS (EET level: viewing, constructing) concentrated more on the solution, spent less time unproductively, and needed less help from the teaching assistant; in addition, they developed better code than pairs of students who used a text editor (EET level: no viewing).

Hübscher-Younger and Narayanan [2003] developed a Web-based system that allowed students to publish their own algorithm representations (text, pictures, animations, multimedia) and discuss them on the Web. They concluded that students who actively participated in this activity achieved higher grades than the passive students who might have only viewed and commented the other students' presentations.

Jehng and Chan [1998] designed and evaluated a distributed visual learning environment for LISP-LOGO that supported collaborative learning. The results showed that students who learned collaboratively, either face-to-face or at a distance, outperformed individual learners in program generation tasks, but that all groups performed equally well in program evaluation and completion tasks. This shows that while collaborative visual learning can be more beneficial compared to individual learning, the improvement can depend on the specific learning task.

Hundhausen [2005] proposed the *communicative dimensions (CD)* framework as a theory on the use of visualizations as communication tools.[2] CD describes the aspects of visualization environments that have an effect on communication between its users in six dimensions: programming salience, provisionality, story content, modifiability, controllability, referencability. While the CD framework is concerned solely with the properties of a visualization tool, the ET framework concerns itself with the interaction between users and visualizations. We think it is very important to understand this interaction and its different levels in order to make tools that support successful collaboration.

2.4 Successful Collaboration Processes

The notion of a successful collaboration process is controversial and not easy to define.

Meier et al. [2007] used a combination of top-down and bottom-up approaches in order to form a description of a successful collaboration. They carried out a comprehensive review of literature, focusing on the aspects of

[2] CD was inspired by the Cognitive Dimensions framework for analyzing individual user interaction with software tools [Green and Petre 1996].

Table II. Aspects and Dimensions of a Successful Collaboration Process [Meier et al. 2007]

Aspect	Dimension
Communication	1) Sustaining mutual understanding
Communication	2) Dialogue management
Joint information processing	3) Information pooling
Joint information processing	4) Reaching consensus
Coordination	5) Task division
Coordination	6) Time management
Coordination	7) Technical coordination
Interpersonal relationship	8) Reciprocal interaction
Motivation	9) Individual task orientation

a successful collaboration (top-down). In addition, they used a data-driven approach to create dimensions from empirical data under each aspect (bottom-up) [Spada et al. 2005]. The study identified five aspects and under them nine dimensions that describe various perspectives of a successful collaboration (see Table II). Furthermore, Meier et al. [2007] used these dimensions in a rating scheme, which they validated with empirical data. The results showed that the high scores on the dimensions of the collaboration process correlate strongly with good results of the collaboration.

The problem with their rating scheme is the difficulty of achieving high inter-rater reliabilities in the ratings. However, this does not mean that the dimensions and aspects of the successful collaboration are not reasonable, and do not reflect the qualities of a successful collaboration. It just means that it is difficult to judge how they appear in the collaboration process.

Teasley [1997] investigated the importance of discussions during collaboration and showed that the amount of discussion is an important part of successful collaboration. However, the amount of transactive reasoning in discussions seems to be an even stronger factor for successful collaboration. Transactive reasoning means talking about one's own thinking process (i.e., reasoning) or one's understanding of the partners' thinking processes [Berkowitz and Gibbs 1983]. In our context, this would mean, for example, that a student talks about what different components of the visualizations mean to him/her or what will happen next in the visualization. Teasley [1997] also showed that a partner is not necessary for transactive reasoning to happen, although the likelihood for it to happen increases when a partner with similar knowledge level is available.

We can summarize these results as showing that successful collaboration requires interaction (dimensions: 1, 2, 4, 8 and Teasley's research), coordination (dimensions: 3, 5, 6, 7) and motivation (dimension: 9).

As discussed in the previous section, Ebel and Ben-Ari [2006] have shown that program visualizations have positive affective effects (i.e., lengthened attention) and anecdotal evidence exists that animation increases students' motivation [Naps et al. 2002]. In addition, Janssen et al. [2007] have showed that augmenting visualizations support the coordination of the collaboration. In this article, we will study how visualization tools affect the interactions in collaboration.

Table III. The Extended Engagement Taxonomy

No viewing (*)	There is no visualization to be viewed but only material in textual format. For example, the students are reviewing the source code without modifying it or they are looking at the learning materials.
Viewing (*)	The visualization is viewed with no interaction. For example, the students are looking at the visualization or the program output.
Controlled viewing	The visualization is viewed and the students control the visualization, for example by selecting objects to inspect or by changing the speed of the animation. This has been deemed important, for instance by Rößling and Naps [2002].
Entering input	The student enters input to a program or parameters to a method before or during their execution.
Responding (*)	The visualization is accompanied by questions which are related to its content.
Changing (*)	Changing of the visualization is allowed during the visualization, for instance, by direct manipulation.
Modifying	Modification of the visualization is carried out before it is viewed, for example, by changing source code or an input set.
Constructing (*)	The visualization is created interactively by the student by construction from components such as text and geometric shapes.
Presenting (*)	Visualizations are presented and explained to others for feedback and discussion.
Reviewing	Visualizations are viewed for the purpose of providing comments, suggestions and feedback on the visualization itself or on the program or algorithm.

3. EXTENDING THE ENGAGEMENT TAXONOMY

3.1 Engagement Levels

The categories of the ET are primarily based on work in AV research, and thus reflect the types of engagement support that are found in AV tools. However, other types of engagement are supported in *PV* tools, and we find it necessary to extend the ET framework in order to capture these differences. Whereas in AV the interaction of the student with the software is more or less restricted to modifications of the visualization itself to the extent allowed by the tool, in PV the opportunities for engagement include both interactive input and, more importantly, the ability to modify the source code that is the basis for the visualization. Consider the software tools we used (see below): In Jeliot, dynamic animations are generated automatically whenever the source code is changed, and BlueJ is based upon interactive calls of methods of a Java class that are regenerated immediately upon modification of a program.

These considerations guided the development of our *extended engagement taxonomy (EET)* shown in Table III. The levels marked with (*) belong to the original ET, although some definitions were slightly modified. Note, in particular, that *changing* in the ET has been divided into two categories, *changing* and *modifying*. We have added new categories: *controlled viewing*, *entering input*, and *reviewing*. Reviewing is different from presenting in that there is not a specific presenter of the visualization; this category (based on a proposal of Oechsle and Morth [2007]) was added for completeness, although it did not occur in our experimental data.

3.2 Linking Engagement to Collaboration Process

Currently, the ET and EET can be used to generate testable hypotheses only about learning outcomes in individual learning. Based on the evidence from Hundhausen and Brown [2008], Myller et al. [2007b] and Laakso et al. [2008], the learning outcome predictions hold also in collaborative learning with visualization. Although the learning results are important, the process leading to them needs to be studied as well, especially in collaborative learning, because in that context, visualization can affect both inter- and intrapersonal learning.

Our goal is to describe how the engagement level affects the (collaborative) learning process. We propose a hypothesis that extends the applicability of EET to collaborative learning with visualizations, with a special emphasis on the collaboration process: the higher the level of engagement between the collaborators and the visualization, the higher the increase in communication and collaboration during the collaborative learning process. This hypothesis builds on the work of Roschelle [1996], Naps et al. [2002], Suthers and Hundhausen [2003], and Hundhausen [2005] as discussed in Section 2, by explicating the connection between engagement and collaboration.

Although there is no previous research that investigated the same hypothesis, there is indirect supporting evidence that was obtained in another study. In the experiment described in Hundhausen and Brown [2008] (see Section 2.3), the collaborative use of visualization at the higher engagement level enhanced the learning process, because pairs of students working with ALVIS held more discussions with each other and less with the teaching assistant (i.e., they needed less help from outside the group), they worked more on the solution, and they had fewer unproductive periods. This provides initial support for our hypothesis by showing in what ways the higher engagement level might enhance collaboration. In order to further evaluate the validity and applicability of the hypothesis, we carried out an empirical study that tests it explicitly.

4. RESEARCH METHODOLOGY

4.1 The Research Setting

In order to verify the hypothesis, we carried out a causal-comparative study in order to understand how the use of visualization tools at different levels of engagement and the collaboration process are correlated. The *causal-comparative method* [Gall et al. 2006] was selected because in the study we are observing both the dependent and independent variables, and could not control the independent variable, because we wanted to maintain the high ecological validity. This method is used when the independent variable cannot be controlled (e.g., independent variable is gender or in our case the observed engagement level of the visualization). A causal-comparative study cannot prove causality, but it can show that correlation between the dependent and independent variables exists. Because our hypothesis is about finding a positive correlation between the engagement levels and students collaborative activities, this is a reasonable methodology to be used in the study.

The study was carried out in an introductory programming course at the University of Joensuu during the autumn of 2005. The course contained 40 hours of lectures (two-hour lectures twice a week for ten weeks), and two-hour recitation sessions every week, where students presented their solutions to the assignments. Every week students took part in a compulsory two-hour computer laboratory session, where they solved exercises with the help of an instructor. Three of the sessions each week were taught by one instructor (*I1*) and two sessions by another (*I2*). One of the instructors was the course lecturer. Neither the lecturer nor the other instructor knew the purpose of the study and were not associated with it in any way.

We investigated the use of the *BlueJ* [Kölling et al. 2003], an educational development environment, and *Jeliot 3* [Moreno et al. 2004], a program animation tool, on different levels of engagement during those laboratory sessions. We were not trying to compare the tools, but rather to analyze how the differing levels of engagement promoted by the tools affected the collaboration.

Initially, we planned to use a between-subject design, and, therefore, one of the sessions of each instructor was randomly selected to belong to a control condition using only BlueJ, while the other three sessions formed the treatment groups using both Jeliot 3 and BlueJ. We needed to abandon this design because a large number of students dropped the course a few weeks before the final exam; as a result, the groups became biased and we could not make a proper comparison of the learning processes and learning outcomes. Thus, we decided to use the the level of engagement as the independent variable—regardless of which tool was being used—and the original division into treatment and control groups became superfluous. We used data only from the treatment groups in order to get data when both Jeliot 3 and BlueJ were used by the same groups.

In order to control for the differences between the two instructors, we analyzed only those groups that were taught by a single instructor (i.e., *I1*). To control for a learning effect (where the behavior of the students would change as they became more familiar with the tools), we analyzed data from a single week near the beginning of the course.

4.2 Participants

There were a total of five sessions of 20 students each participating in the compulsory laboratory sessions weekly. The total number of the students who gave their consent to participate in the research was 89. Those who did not agree to participate in the study worked in small groups where no data collection was carried out.

The programming course was primarily taught to first year computer science majors. However, a significant proportion (about 60%) of the students taking the course were students majoring in other subjects who studied computer science as a minor. Additionally, there were major students from previous years who had not yet passed the course. Since the current study is not a controlled experiment, randomization is not necessary. We excluded two students who had very extensive programming experience. Otherwise, a post-hoc analysis showed that the remaining students had similar background knowledge and relatively little previous experience in programming.

Because we only analyzed sessions taught by instructor (*I1*) and excluded the original control session, we, altogether, analyzed data from two sessions, otherwise, ten groups of 3–4 students each, a total of 39 students.

4.3 Materials

During the laboratory sessions students were presented with exercises related to the topics of the course. There were no mini-lectures in the beginning of the session, but the exercises were related to the lectures that were given during the same or previous week to the whole course. The teacher handed out the exercises at the beginning of the session and circled around to help out the students. Only if the instructor spotted that several groups had the same misconception or were stuck on the same issue, the teacher went to the front of the class and announced the issue to all the groups and briefly explained it. When the students thought that they were ready with an exercise, they summoned the teacher to check it.

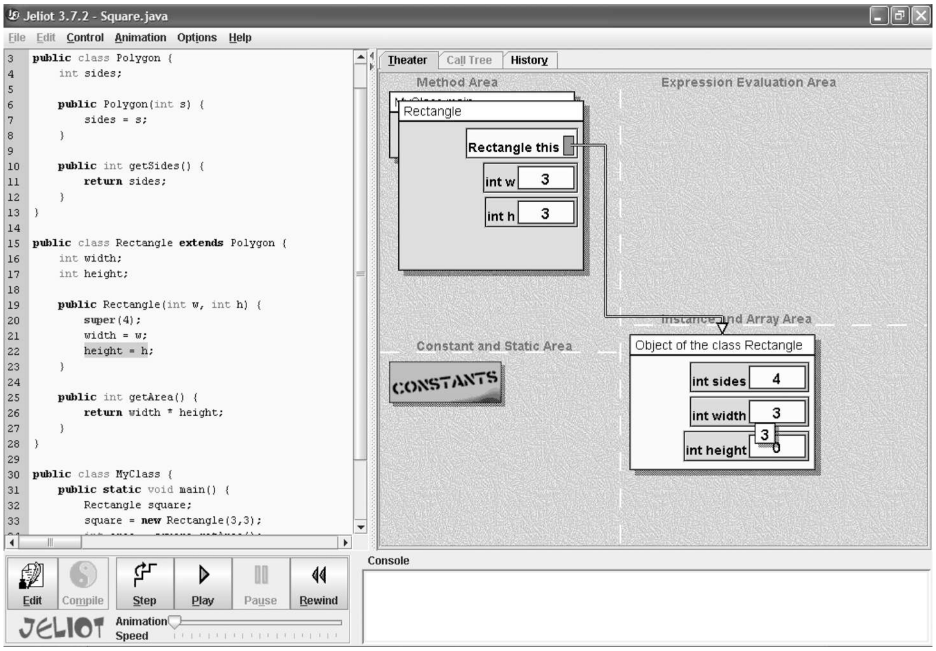
The exercises varied from program construction and modification to debugging. For example, students were given a program code and told what it was supposed to do and they needed to check if the program did what was expected and if it did not, they needed to correct the program. In another exercise, the students were given a skeleton of a program and asked to fill in the missing parts or to create an accompanying class that enabled the program to work as expected. The exercises that were solved in groups were purposefully more difficult than the ones solved individually, because pilot studies indicated that if one of the students could solve the exercise independently, there was neither collaboration nor communication between the students.

4.4 Visualization Tools

Students used both BlueJ and Jeliot during the laboratory sessions. Both tools have proved to be effective in improving the learning of elementary computer science and programming [Ben-Bassat Levy et al. 2003; Haaster and Hagan 2004; Ragonis and Ben-Ari 2005].

The user interface of Jeliot 3 is illustrated in Figure 1. The source code editor is in the left-hand pane, while the right-hand pane is used to display the visualization. VCR-like buttons to control the visualization are located in the lower left corner. Fully dynamic animation of the data and control flow of the program is displayed, including method calls, object construction, and expression evaluation. The animation is created automatically from the source code, so that the student needs only to learn to use the control buttons for the visualization.

Figure 2 shows the user interface of BlueJ. The class diagram is shown in the middle of the window. The student can interactively instantiate an object of a class by right-clicking on the class and selecting the constructor from a popup


Jeliot 3.7.2 - Square.java
File Edit Control Animation Options Help
Theater Call Tree History
Method Area
Expression Evaluation Area
Rectangle
Rectangle this
int w 3
int h 3
Instance and Array Area
Constant and Static Area
CONSTANTS
Object of the class Rectangle
int sides 4
int width 3
int height 0
3
Console
Edit Compile Step Play Pause Rewind
JELIOT
Animation Speed

```
public class Polygon {
    int sides;

    public Polygon(int s) {
        sides = s;
    }

    public int getSides() {
        return sides;
    }
}

public class Rectangle extends Polygon {
    int width;
    int height;

    public Rectangle(int w, int h) {
        super(4);
        width = w;
        height = h;
    }

    public int getArea() {
        return width * height;
    }
}

public class MyClass {
    public static void main() {
        Rectangle square;
        square = new Rectangle(3,3);
```

Fig. 1. User interface of Jeliot 3.

menu. The objects are then shown at the bottom of the screen as red icons, and the methods of an object can be interactively invoked by right-clicking and selecting the method from a pop-up menu. Parameters of constructors and method calls are also entered interactively.

In the experiment, Jeliot was used as a BlueJ plugin [Myller et al. 2007a] that performs code synchronization between the tools. This allowed students to freely switch between the two tools.

BlueJ and Jeliot were both introduced in the laboratory session of the first week. Students used them individually during the first session so that they were familiar with the tools before the tools were used collaboratively.

4.5 Procedure

There were a total of ten laboratory sessions for each class; during four of them—the 2nd, 4th, 6th, and 8th—students worked in small groups. The students were randomly assigned (by a computer program) to small groups of three or four students, and the membership in the groups was unchanged throughout the course. Due to the high drop-out rate in the course, the number of students in some groups became too low and the groups needed to be merged in the second half of the course. Thus, the materials from the week eight are not comparable to the materials of the other weeks because the groups are not the same.

The learning process in the groups was filmed with video cameras, one for each small group. The filming was done by the first author who was not associ-

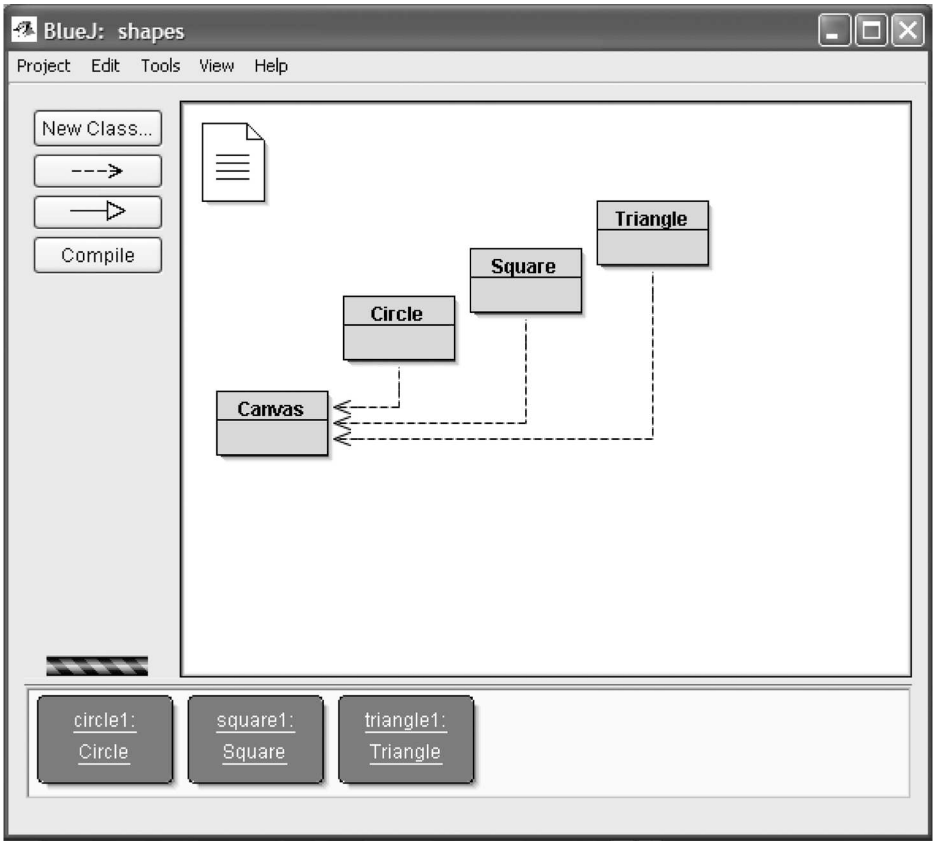


Fig. 2. User interface of BlueJ.

ated with the course. The camera was positioned so that students' movements, facial expressions, and the screen could be recorded. Although minimally invasive, this setting posed some problems in the recordings, because students moved during the lesson and the video cameras could not be always adjusted so that all students' faces were recorded.

Because BlueJ was the primary tool of the course both during the lectures and during the laboratory sessions, the instructor was advised to encourage the students to use Jeliot so that both tools would be used evenly. Since the instructor used both tools as necessary in different exercises, the students were exposed to both tools and could autonomously make decisions when to use each of the tools.

4.6 Data Analysis

To analyze how visualization affected the students' learning processes, we repeatedly viewed those parts of the video materials that contained episodes relating to the execution of a program using one of the tools. These episodes were characterized by extensive use of the visualization capabilities of the tools. Because of the large amount of video-taped material (1.5 hours per small group per week), we randomly sampled 30 two-minute-long episodes. A similar number of episodes were sampled from each group, as were a similar number of episodes using each of the tools. As mentioned above, the analysis was restricted to the sessions of the second week.

In the analysis of the data, we classified each five-second segment of the video according to several classification schemes in order to analyze the level of engagement, and the behavior and the discussions of the students. We used the extended engagement taxonomy (EET) as the independent variable to capture the changes in the level of engagement during an episode in which collaborative learning took place. The EET levels were presented in Table III. Each segment was classified as belonging to a single EET level; in the case where several engagement levels were applicable, the segment was classified at the level that lasted the longest time, or if the times were the same, at the level that was higher in the EET.

The dependent variables were aspects of the students' behavior and discussions; we classified the activities (Section 4.6.1), the discussion contents (Section 4.6.2) and transactive reasoning (Section 4.6.3). The differences found in these classifications were analyzed using the statistical tests Cohen's κ and χ^2 test of independence. To adhere to the limitation of χ^2 test the categories of data with counts less than ten were excluded from the analysis.

4.6.1 *Activities.* We adopted and modified the activity categories of Hundhausen and Brown [2008] to classify the five-second clips of the episodes (see Table IV). We were interested in episodes, during which the students in the group discussed aspects of the exercise (the program, the Java language, or the visualization), and we categorized these episodes according to the activities that accompany the conversation. We removed some of the categories (such as *executing code* and *working on solution*) that were irrelevant for the analysis since we chose to analyze episodes where students were, in fact, carrying out these activities. In addition, we added new categories related to different types of conversing (i.e., conversing with gesturing and conversing with drawing), and changed the priorities of the categories in order to prioritize discussions. Thus, if any kind of discussion happened, it was then assigned to one of the discussion categories. More specialized discussions were given a higher priority. We wanted to distinguish between discussions within the group and discussions with the teacher, so conversing to the instructor had higher priority so it could be distinguished from other types of discussions that happened within the group.

Each five-second period of the video was assigned to a single category based on the activities of the groups. Because the members of a group might perform several activities simultaneously, the priorities given for each category were used to resolve the ambiguities. An episode was assigned the category with the highest priority (lower numbers mean higher priority). Furthermore, we assigned the number of participants for each activity in order to determine if certain EET levels increase or reduce the participation of students.

Table IV. Activity Categories

Priority	Category	Description
1	1 Conversing to an instructor	The students discuss the exercise with the instructor.
2	2 Conversing with gesturing	The students perform gestures such as pointing at the screen when discussing the exercise.
3	3 Conversing with drawing	The students discuss the exercise with the help of drawing.
4	4 Conversing	The students discuss the exercise (without any of the above activities).
5	5 Listening to an instructor	The students are listening to the instructor who is talking with the group or announcing something to the whole class.
5	6 Looking at or searching for course materials, Internet resources or example.	Self-explanatory.
5	7 Reviewing the exercise	The students are looking at the online or hard-copy description of the problem.
5	8 Reviewing error messages	The students are reviewing error messages produced by the environment.
6	9 Silent work	The students are working silently; for example, looking at the visualization or entering input without comment.
8	10 Other	No observable activity or the activity is off-task.
9	-1 Indeterminable	The event cannot be categorized, for example, because of a technical problem in the recording.

4.6.2 *Discussion Contents.* We wished to investigate the actual content of the students' discussions in order to determine if the engagement level had any impact on the contents. If an activity was classified as (i) any type of conversing, (ii) other but it contained talking, or (iii) undetermined but it contained talking, it was further classified into one of ten discussion content categories (see Table V). The categories were exclusive, and the decision to classify an event into a category was based upon the discussions during the five-second period. If several categories were applicable, we used the category with the highest priority, and if there were several with the same priority then the one that happened first was chosen. The categories in Table V were adapted from Hundhausen and Brown [2008]. As that study dealt with work at the algorithmic level, we changed the relevant content categories so that they refer to programs instead of algorithms. Furthermore, we removed categories that were related to the ALVIS animation system that we did not use, and we added a category, *programming concepts*.

4.6.3 *Transactive Reasoning.* Since transactive reasoning has been found to have a positive impact on learning outcomes [Teasley 1997], we measured the amount of transactive reasoning in order to see if it can help determine how visualization affects the collaboration process.

If an activity was classified as any type of conversing, as other but it contained talking, or as undetermined but it contained talking, it was further clas-

Table V. Discussion Content Categories

Priority	Category	Description
1	1 Program	The content relates to the program code: "This line contains a while-loop."
1	2 Program behavior	The content relates to the program's behavior: "Now it repeats 10 times."
1	3 Programming concepts	The content relates to programming concepts in general not directly related to the current program: "What does `double` mean?"
1	4 Tool	The content relates to the tools currently being used tool: "How can I display the value of a variable?"
1	5 Error detection	The content relates to programming errors and their detection: "I spotted an error!"
1	6 Error correction	The content relates to the correction of an error: "If we change the value of this variable, that will solve the problem."
1	7 Visualization	The content relates to understanding the visualization itself: "How is this box related to the program?"
2	8 On-topic (other)	The content relates to the current task but cannot be placed into one of the previous categories.
2	9 Off-topic	Self-explanatory.
3	10 Indeterminable	The content cannot be categorized, for example, because of a technical problem in the recording.
0	-1 Not applicable	There is no content in this activity; for example, there is no talking in the segment.

sified into one of twelve transactive reasoning categories (see Table VI). These categories were exclusive. The decision to classify an event was based on the discussions and activities during a five-second block of the video. If several categories were applicable, we used the category with the highest priority, and if there were several categories with the same priority then the one that happened first was chosen. The categories in Table VI were adapted from Teasley [1997] and Berkowitz and Gibbs [1983]. We added the prediction category used by Teasley [1997] to the categories of Berkowitz and Gibbs [1983].

5. RESULTS

5.1 Inter-Rater Reliability

In order to test the reliability of the classification schemes used in the study, a set of ten episodes (a total of 240 five-second blocks) were classified by two raters, the first author, who classified all the video material used in the study, and the second author, who classified only the set of ten episodes.

In a pilot, both raters analyzed three other episodes in order to reach agreement on how to classify the observed behavior. The classification schemes and the coding manual were updated as a result of the discussions leading to consensus between the raters.

The inter-rater reliabilities are presented in Table VII. All Cohen's κ values indicated that it is very unlikely ($p < 0.001$) that this level of agreement is achieved by chance. Furthermore, the *EET* and Activities classification has a

Table VI. Transactive Reasoning Categories

Priority	Category	Description
1	0 Prediction	A student tries to predict what will happen next and justifies the prediction.
1	1 Feedback Request	A student ensures that others understand or agree with his/her position
1	2 Paraphrase	A student paraphrases a discourse of another student in order to demonstrate that he/she understands it.
1	3 Justification	A student justifies his/her position or reasoning.
1	4 Juxtaposition	A student explains the differences between the positions or reasoning of other students and his/her own.
1	5 Completion	A student completes another student's reasoning, for example, by filling out an unfinished sentence.
1	6 Clarification	A student explains his/her reasoning in order to ensure that others understand it.
1	7 Refinement	A student elaborates or qualifies his/her position in order to to defend against criticism.
1	8 Extension	A student elaborates on a previous discourse.
1	9 Criticism	A student criticizes the reasoning or position of another student and explains the reason for the criticism.
1	10 Integration	A student combines different views into one common statement.
2	11 No transactive reasoning	The discussion contains no transactive reasoning.
0	-1 Not applicable	This categorization is not applicable; for example, there is no talking in the segment.

Table VII. Inter-rater Reliabilities. * $p < 0.001$

Classification Scheme	Agreement Percentage	Cohen's κ
EET	76.3%	(0.66 *)
Activities	76.3%	(0.68 *)
Number of Participants	58.3%	(0.42 *)

substantial agreement and the Number of Participants has moderate agreement based on classification of Cohen's κ-measures given by Landis and Koch [1977].

For the two other categorizations (Discussion Contents and Transactive Reasoning) inter-rater reliabilities were low. We believe that this was due to several factors: 1) subtle cues of the discussion contents or transactive reasoning (e.g., only one word could indicate if students discussed about program, its behavior or programming concept) which might have been misinterpreted by the raters, and 2) noise in the natural environment (i.e., classroom with several groups) made it sometimes difficult to interpret exactly what the group was discussing. We still include in the article results of the discussion contents and transactive reasoning. Although they cannot be used as definitive evidence supporting our hypothesis, the results do indicate that these categorizations are consistent with the previous categorizations.

Table VIII. The Distribution of Activities on Different EET Levels

	Conversing w/I	Conversing w/G	Conversing	Listen t/I	Review Ex	Review EM	Silent	Count
No viewing	6.8%	12.4%	27.8%	27.2%	0.0%	1.2%	24.7%	162
Viewing	4.4%	6.1%	25.4%	15.7%	0.9%	0.0%	47.5%	343
Controlled viewing	8.6%	0.0%	22.9%	34.3%	0.0%	0.0%	34.3%	35
Entering input	5.5%	8.0%	42.9%	12.3%	0.0%	1.2%	30.1%	163
Changing	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	2
Modifying	0.0%	16.7%	50.0%	0.0%	0.0%	0.0%	33.3%	6
Constructing	22.2%	0.0%	0.0%	77.8%	0.0%	0.0%	0.0%	9
Overall	5.6%	7.6%	29.6%	19.3%	0.4%	0.6%	36.9%	720

Legend: w/I = with instructor; w/G = with gestures; t/I = to instructor; Ex = exerices; EM=error messages

5.2 Activities

The distribution of the activities on each EET level is presented in Table VIII. EET levels that contained fewer than ten observations or that had categories that did not contain any observations (controlled viewing, changing, modifying and constructing) were excluded from the analysis. Activity columns that contained observations only on one or two EET levels (i.e., looking at or searching for examples or course materials, and reviewing the exercise) were also excluded due to the restrictions of χ^2-test. These categories only contributed less than eight percent of the overall data.

The distributions of the EET levels no viewing, viewing and entering input were compared, first collectively ($\chi^2(8) = 48.5, p < .01$), and then pairwise (no viewing vs. viewing $\chi^2(4) = 28.4, p < .01$; no viewing vs. entering input $\chi^2(4) = 17.0, p < .01$; viewing vs. entering input $\chi^2(4) = 20.9, p < .01$). All tests were found to be statistically significant, meaning that the EET level has an effect on the distribution.

Figure 3 shows the distributions of activities for the three most common EET levels collapsed into three columns. The different forms of conversing are combined into one, *sum of conversing*, and the categories that did not have observations for all engagement levels were removed as they contributed less than eight percent of the data. The distributions were first compared collectively ($\chi^2(4) = 41.6, p < .01$), and then pairwise (no viewing vs. viewing $\chi^2(2) = 25.0, p < .01$; no viewing vs. entering input $\chi^2(2) = 37.6, p < .01$; viewing vs. entering input $\chi^2(2) = 28.2, p < .01$). All tests were found to be significant. Figure 3 shows that entering input produced the greatest amount of conversation. When students were not viewing a visualization, they listened to the teacher more often than on any other EET level. When students were viewing a visualization they were more often silent.

Figure 4 illustrates the difference between conversing and silent activities performed by the groups when either on viewing or on entering input level. On entering input level over half of the activities contained discussions whereas in viewing level the percentage was approximately 35%. Almost the opposite happens with the amount of silence.

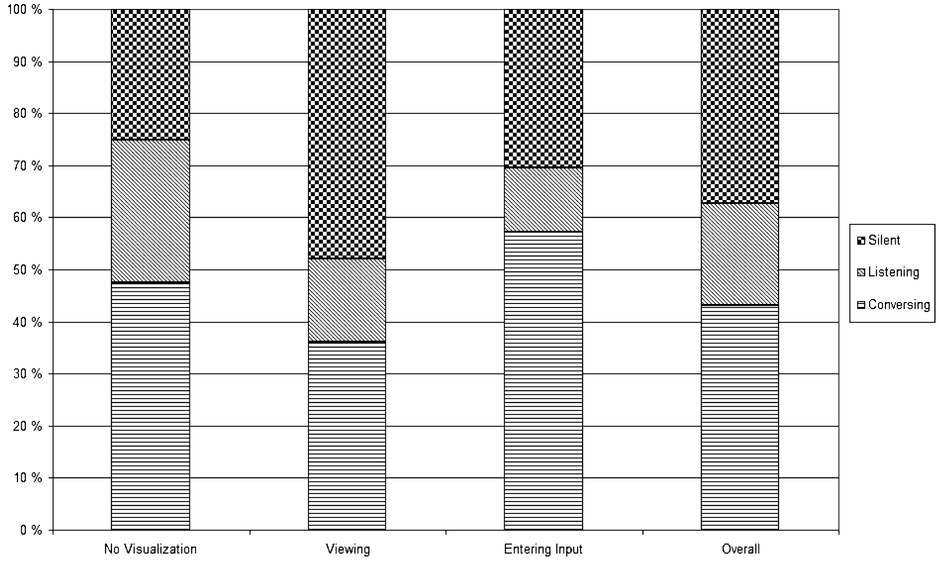


Fig. 3. Activity distributions on different EET levels.

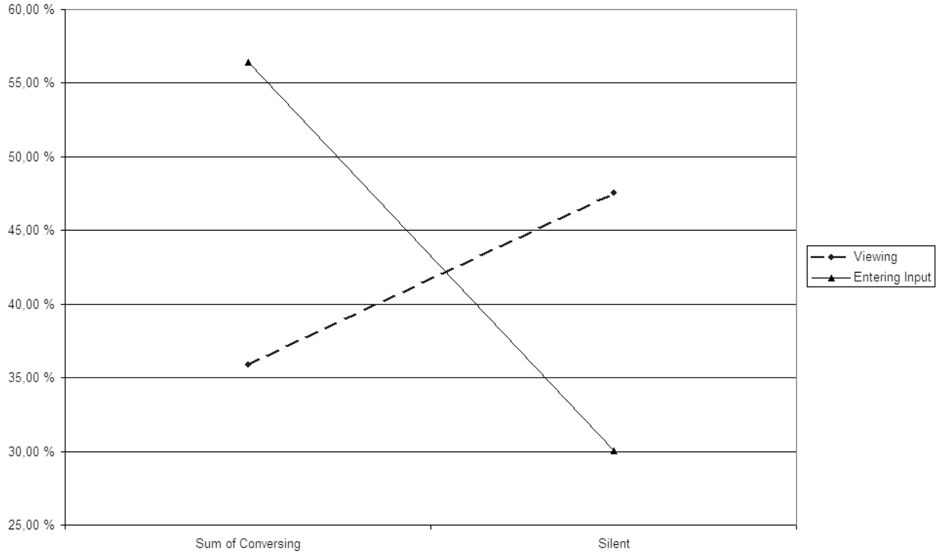


Fig. 4. Viewing and entering input EET-levels compared on sum of conversing and silent categories.

Table IX shows how the distribution of activities differs on each EET level, depending on the tool used by the students. Jeliot provides support only for the first four EET levels and the modifying level, which did not appear in the data, therefore, there are no results for the higher levels. The within-tool distributions of the activities for the most common EET levels (no viewing, viewing and entering input) were compared to each other and differences

Table IX. Distribution of Activities on Each EET Level and When Using a Particular Tool. (Legend as in Table VIII)

	Conversing w/I	Conversing w/G	Conversing	Listen t/I	Silent	Count
Jeliot						
No viewing	4.8%	16.7%	23.8%	33.3%	21.4%	42
Viewing	0.4%	4.9%	22.5%	12.7%	59.6%	245
Controlled viewing	8.3%	0.0%	41.7%	16.7%	33.3%	12
Entering input	1.8%	10.9%	25.5%	10.9%	50.9%	55
BlueJ						
No viewing	7.6%	11.0%	29.7%	25.4%	26.3%	118
Viewing	14.7%	9.5%	33.7%	24.2%	17.9%	95
Controlled viewing	8.7%	0.0%	13.0%	43.5%	34.8%	23
Entering input	7.6%	6.6%	52.8%	13.2%	19.8%	106
Changing	0.0%	0.0%	0.0%	100.0%	0.0%	2
Modifying	0.0%	16.7%	50.0%	0.0%	33.3%	6
Constructing	22.2%	0.0%	0.0%	77.8%	0.0%	9

Table X. The Average Number of Participants at Each EET Level

EET	Average Number of Participants
No viewing	3.1
Viewing	3.1
Controlled viewing	3.5
Entering input	2.9
Changing	4.0
Modifying	3.0
Constructing	3.2
Overall	3.1

were found to be statistically significant (Jeliot: $\chi^2(8) = 35.0, p < .01$; BlueJ: $\chi^2(8) = 19.9, p < .05$). The activity distributions on both tools are similar to the overall distribution shown in the Figure 3 (cf., previous paragraph).

5.3 Participation

Table X shows the average number of participants on each EET level. In the most frequently occurring levels (no viewing, viewing, entering input), the number of participants is almost the same. There were no statistically significant differences and the number of participants on each activity is large enough to argue that the students were truly collaborating.

5.4 Discussion Contents

All the activities that contained discussions were further classified based on their discussion contents. Table XI shows the distribution of the frequency of topics during the conversations when the students were working on one of the EET levels.

The variability of the discussion contents between EET levels was large; for example, on the no viewing level, the program category contains 25.7% of the data, while there are no data for program category on the *viewing* level.

Table XI. Discussion Topics Distribution on Each EET Level

EET /	P	PB	PC	Tool	ED	EC	Visual-	On-	Off-	I	Count
Content							ization	topic	topic		
No viewing	25.7%	24.9%	0.0%	9.9%	4.0%	9.9%	0.0%	22.8%	2.0%	1.0%	101
Viewing	0.0%	33.1%	2.1%	14.8%	9.2%	2.8%	3.5%	29.6%	2.1%	2.8%	142
Controlled viewing	0.0%	9.5%	9.5%	42.9%	4.8%	0.0%	0.0%	33.3%	0.0%	0.0%	21
Entering input	2.1%	37.9%	4.2%	11.6%	4.2%	1.1%	1.1%	34.7%	0.0%	3.2%	95
Changing	0.0%	0.0%	0.0%	50.0%	0.0%	0.0%	0.0%	50.0%	0.0%	0.0%	2
Modifying	0.0%	25.0%	0.0%	0.0%	50.0%	25.0%	0.0%	0.0%	0.0%	0.0%	4
Constructing	33.3%	0.0%	33.3%	33.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	3

Legend: P=Program; PB=Program behavior; PC=Program concepts; ED=Error detection; EC=Error correction; I=Indeterminable

Table XII. Transactive Reasoning on Each EET Level

EET	Transactive reasoning	Count
No viewing	5.0%	101
Viewing	3.6%	139
Controlled viewing	0.0%	21
Entering input	6.5%	93
Changing	0.0%	2
Modifying	33.3%	3
Constructing	0.0%	3

Therefore, the χ^2 test might not be reliable. The three most frequent EET levels (no viewing, viewing and entering input) were included into the analysis. The test was first done for all three levels showing that the discussion content distributions and the EET levels are related ($\chi^2(18) = 85.1, p < .01$). However, pairwise comparison revealed that only differences between no viewing and viewing ($\chi^2(9) = 54.1, p < .01$), and between no viewing and entering input ($\chi^2(9) = 39.6, p < .01$) were significant, meaning that those levels have different discussion content distributions. This also means that the distributions of the discussion contents were very similar on both of the EET levels, viewing and entering input, based on the statistical tests.

5.5 Transactive Reasoning

Less than five percent of all the discussions contained transactive reasoning. Therefore, we do not report the results for each type separately, but, rather, in Table XII we show the percentages of discussions that contained and did not contain transactive reasoning. From the most frequent EET levels, entering input had the highest percentage of transactive reasoning; however, the differences were not statistically significant.

6. DISCUSSION

The study partially confirms our hypothesis that the level of engagement on which students select to work with the visualization tool affects the quality of collaboration; that is, the engagement level and the interaction between the students are correlated. This seems to be true especially when a tool supports engagement on the levels viewing and entering input: the latter increases the amount of discussion significantly and reduces the time when students are silent, the former does the opposite. The level entering input also increases the amount of transactive reasoning, although the differences are not statistically significant. The increase in the amount of discussion changes neither the contents of the discussions nor the participation of the students, and similar numbers of students discuss similar topics on both levels, viewing and entering input.

To summarize, this study shows that EET levels are positively correlated with the amount of interaction, and it is interaction which is an important component of the successful collaboration (Section 2.4). However, this does not change the contents of the interactions and actually increases, although not significantly, the amount of transactive reasoning, which is positively correlated with learning outcomes.

At the lowest level of engagement (i.e., no viewing), the hypothesis does not hold and the quality of the collaboration lies between that of entering input and viewing. It seems that the no viewing level promotes different kinds of interactions and communication among the students than the other levels. Students' discussions are more about the program code, which is natural, because the program code is often the only representation available at this level. Furthermore, students listen to and discuss with the instructor more often, which can mean either that they are seeking help or that they are listening to instructor's guidance more often when they are working on this level or that the teacher guides them to this level when he/she comes to help the students. It is reasonable to assume that this compensates for the lack of help and guidance that is available from the visualization tool. This result partially replicates the results of Hundhausen and Brown [2008] discussed in the Section 2.3.

The average number of participants taking part in the collaboration is roughly the same at all levels of engagement. Three students seem to be optimal for collaborating on a single computer that is running a visualization tool. We observed from the videos that in many cases when there were four students, one of them sat far from the computer and was passive. Nevertheless, for some groups the amount of collaboration among all four students was high, a result we attribute to their interpersonal skills.

Although a comparison of Jeliot and BlueJ is not within the scope of this article, a preliminary analysis indicates that it was the level of engagement that was more influential than the specific tool. A currently unpublished follow-up analysis of video protocols [Korhonen et al. 2008] from a study, in which students were using the TRAKLA2 system to learn data structures in small groups (see the report of the original study in Myller et al. [2007b] and in Laakso et al. [2008] on students' learning results), found the same correlation between engagement levels and the amount of communication and collaboration. This also supports the hypothesis and shows that it is independent of the tool that is being used.

Our experiment was carried out in an actual classroom and it has high ecological validity, but this also means that we could not control all variables as we could have done in an experimental setting. Therefore, we could not establish that the engagement level caused the increase in the interactions and the quality of collaboration. We can only say that they are correlated, so the causation, if any, could be in either direction. On the one hand, there are certain levels of engagement that were not always controlled by the students, but happened as a side effect of the animation or other actions. For example, when students were viewing an animation, students needed to enter input whenever the visualized program needs input. Thus, they engaged on entering input, although they only selected the viewing of the animation. On the other hand, an activity could change the engagement level; for example, when students were discussing with the instructor, s/he proposed the use of a visualization on a certain engagement level (e.g., seeing the source code instead of the animation).

In this study, we have not taken into consideration how the correlation between EET levels and collaboration and communication develops over time. Thus, based on the current research, we can definitely describe the correlation only for the first week of the first course on programming. In future research, we plan to study the effects of the use of visualizations on collaboration by analyzing the video materials from several weeks of the course.

6.1 Threats to Validity and Reliability

Although we have tried to make sure that all the steps of the research process would maintain the internal validity and reliability, there are issues that might have affected the results. We have aimed to have a high ecological validity, so we carried out the study in a real classroom environment. We could not randomize the students into the different sessions; however, we randomized the students within the sessions into small groups and by using a post-hoc analysis, we checked that the background variables of the students were similar between the sessions. Thus, we expect that the collected data from several groups is comparable and aggregable.

When the video data was sampled, we selected episodes in the materials from one week. There could be bias in the random selection of the episodes that could affect the results. We tried to minimize the bias by selecting a similar number of episodes from each group for each tool. This should have balanced out the influence of a single group to the final results.

The videos were coded using different coding schemes. The coding schemes were checked for reliability by comparing the results of two coders on a subset of the data. It was found that most of the coding schemes are reliable except for the discussion contents and transactive reasoning categories. Thus, we use them as secondary evidence not as the primary evidence for our hypothesis. The reliability of this secondary evidence needs to be further tested in the future studies.

Some of the EET levels were not completely assessed because we did not have enough data from those levels. Given the high inter-rater reliabilities in classifying all the EET levels, we believe that additional data collection will enable us to analyze other levels in the future experiments.

7. CONCLUSION

We have presented an extension to the Engagement Taxonomy, the Extended Engagement Taxonomy, and used it to investigate the mutual relationship between visualization tools and collaborative learning. Our empirical study demonstrated support for the hypothesis that increasing the level of engagement between learners and the visualization tool results in a higher positive impact of the visualization on the collaboration process.

The EET can be used during the design and development of visualization tools for collaborative learning. There are several design implications that are already implemented in practice:

- We have added a capability for automatic question generation to Jeliot 3 in order to support the EET level responding [Myller 2007].
- We believe that closer linking of BlueJ and Jeliot [Myller et al. 2007a] can increase the engagement level of students. Students should be able to see both BlueJ's object bench and Jeliot's animation at the same time, so that any modification of an object results in a change in the animation. This would allow a step-by-step construction of dynamic visualizations by the students.
- We have combined Jeliot 3 with Woven Stories 2, a collaborative authoring tool [Myller and Nuutinen 2006], in order to provide Jeliot 3 with collaborative editing support and augmenting visualizations that can support online collaborative learning.

There are also pedagogical implications of the findings:

- Higher levels of engagement provide more support for collaborative activities, so instructors should find ways to use visualization tools at these levels.
- The viewing level seems to reduce collaboration significantly because students become passive. Thus, engagement at this level should be avoided, if possible, or it should combined with explanations by the teacher, as was done by Ben-Bassat Levy et al. [2003] with positive results.
- The lowest level of engagement (no viewing) does not decrease interaction and collaboration of students as much as viewing. However, when this level of the EET is used in teaching, the instructor should be aware of the change in the focus of the discussions (the program code) and of the need to provide students with more help and guidance.

Our study suggests new directions in research on engagement in collaborative learning with visualizations. The first step should be an expansion of the analysis of the differences at a finer level of detail; for example, the contents of the discussions between the students could be further analyzed to determine what communicative resources they are referring to during the discussions (e.g., the visualization or the source code). Also, the interaction of learning and time-on-task should be analyzed in order to better understand how the role of the visualization during the students' learning process changes in long run.

Therefore, a longitudinal analysis of the collected data from this study (video materials from sessions during several weeks) could be used to analyze the effects of time and learning, for example, by using a time-series analysis.

We also plan to evaluate the question generation support added to Jeliot 3 and the closer linking of BlueJ and Jeliot. This should shed more light on the effects of the higher levels of the EET.

We think that it is important to create visualization tools that support several engagement levels—especially the higher ones—and make the instructors aware of the potentials that the higher engagement levels can provide so that students and their instructors can use the tools to their benefit during the collaborative learning.

ACKNOWLEDGMENTS

The authors would like to thank Andrés Moreno for the help during the data collection process.

REFERENCES

BEN-BASSAT LEVY, R., BEN-ARI, M., AND URONEN, P. A. 2003. The Jeliot 2000 program animation system. *Comput. Ed. 40*, 1, 15–21.

BERKOWITZ, M. W. AND GIBBS, J. C. 1983. Measuring the development of features in moral discussion. *Merill-Palmer Quar. 29*, 399–410.

BRYANT, S., ROMERO, P., AND DU BOULAY, B. 2005. Pair programming and the reappropriation of individual tools for collaborative programming. In *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work (SIGGROUP'05)*, M. Pendergast, K. Schmidt, G. Mark, and M. Ackerman Eds. ACM Press, 332–333.

EBEL, G. AND BEN-ARI, M. 2006. Affective effects of program visualization. In *Proceedings of the 2nd International Computing Education Research Workshop (ICER'06)*. ACM Press, 1-5.

EVANS, C. AND GIBBONS, N. J. 2007. The interactivity effect in multimedia learning. *Comput. Ed. 49*, 4, 1147–1160.

GALL, M. D., GALL, J. P., AND BORG, W. R. 2006. *Educational Research: An Introduction 8th Ed.* Allyn & Bacon.

GREEN, T. R. G. AND PETRE, M. 1996. Usability analysis of visual programming environments: A "cognitive dimensions" framework. *J. Vis. Lang. Comput. 7*, 131–174.

GRISSOM, S., MCNALLY, M., AND NAPS, T. L. 2003. Algorithm visualization in CS education: Comparing levels of student engagement. In *Proceedings of the 1st ACM Symposium on Software Visualization (SOFTVIS'03)*. ACM Press, 87–94.

HAASTER, K. V. AND HAGAN, D. 2004. Teaching and learning with BlueJ: An evaluation of a pedagogical tool. In *Proceedings of Informing Science + IT Education Conference (InSITE'04)*. Informing Science Institute, 455–470.

HÜBSCHER-YOUNGER, T. AND NARAYANAN, N. H. 2003. Constructive and collaborative learning of algorithms. *SIGCSE Bull. 35*, 1, 6–10.

HUNDHAUSEN, C. D. 2002. Integrating algorithm visualization technology into an undergraduate algorithms course: Ethnographic studies of a social constructivist approach. *Comput. Ed. 39*, 3, 237–260.

HUNDHAUSEN, C. D. 2005. Using end-user visualization environments to mediate conversations: A "communicative dimensions" framework. *J. Vis. Lang. Comput. 16*, 3, 153–185.

HUNDHAUSEN, C. D. AND BROWN, J. L. 2005. Personalizing and discussing algorithms within CS1 studio experiences: An observational study. In *Proceedings of the International Workshop on Computing Education Research (ICER'05)*. ACM Press, 45–56.

HUNDHAUSEN, C. D. AND BROWN, J. L. 2007. What you see is what you code: A “live” algorithm development and visualization environment for novice learners. *J. Vis. Lang. Comput. 18*, 1, 22–47.

HUNDHAUSEN, C. D. AND BROWN, J. L. 2008. Designing, visualizing, and discussing algorithms within a CS 1 studio experience: An empirical study. *Comput. Ed. 50*, 1, 301–326.

HUNDHAUSEN, C. D., DOUGLAS, S. A., AND STASKO, J. T. 2002. A meta-study of algorithm visualization effectiveness. *J. Vis. Lang. Comput. 13*, 3, 259–290.

JANSSEN, J., ERKENS, G., KANSELAAR, G., AND JASPERS, J. 2007. Visualization of participation: Does it contribute to successful computer-supported collaborative learning? *Comput. Ed. 49*, 4, 1037–1065.

JEHNG, J.-C. J. AND CHAN, T.-W. 1998. Designing computer support for collaborative visual learning in the domain of computer programming. *Comput. Hum. Behav. 14*, 3, 429–448.

KÖLLING, M., QUIG, B., PATTERSON, A., AND ROSENBERG, J. 2003. The BlueJ system and its pedagogy. *Comput. Science Ed. 13*, 4, 249–268.

KORHONEN, A., LAAKSO, M., AND MYLLER, N. 2008. How does algorithm visualization affect collaboration? Video analysis of engagement and discussions. *5th International Conference on Web Information Systems and Technologies (WEBIST’09)*. Submitted.

LAAKSO, M.-J., MYLLER, N., AND KORHONEN, A. 2008. Analyzing the extended engagement taxonomy in collaborative algorithm visualization. *J. Ed. Technol. Soc*. To appear.

LANDIS, J. R. AND KOCH, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics 33*, 159–174.

MCDOWELL, C., WERNER, L., BULLOCK, H. E., AND FERNALD, J. 2006. Pair programming improves student retention, confidence, and program quality. *Comm. ACM 49*, 8, 90–95.

MEIER, A., SPADA, H., AND RUMMEL, N. 2007. A rating scheme for assessing the quality of computer-supported collaboration processes. *International J. Comput. Support. Collab. Learn. 2*, 1, 63–86.

MORENO, A., MYLLER, N., SUTINEN, E., AND BEN-ARI, M. 2004. Visualizing program with Jeliot 3. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI’04)*. ACM Press, 373–380.

MYLLER, N. 2007. Automatic generation of prediction questions during program visualization. *Electron. Notes Theor. Comput. Sci. 178*, 43–49.

MYLLER, N., BEDNARIK, R., AND MORENO, A. 2007a. Integrating dynamic program visualization into BlueJ: The Jeliot 3 extension. In *Proceedings of the 7th IEEE International Conference on Advanced Learning Technologies*, J. M. Spector, D. G. Sampson, T. Okamoto, Kinshuk, S. A. Cerri, M. Ueno, and A. Kashihara Eds. IEEE Computer Society, 505–506.

MYLLER, N., LAAKSO, M., AND KORHONEN, A. 2007b. Analyzing engagement taxonomy in collaborative algorithm visualization. In *Proceedings of the 12th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education (ITiCSE’07)*, J. Hughes, D. R. Peiris, and P. T. Tymann Eds. ACM Press, 251–255.

MYLLER, N. AND NUUTINEN, J. 2006. JeCo: Combining program visualization and story weaving. *Informatics Ed. 5*, 2, 267–276.

NAGAPPAN, N., WILLIAMS, L., FERZLI, M., WIEBE, E., YANG, K., MILLER, C., AND BALIK, S. 2003. Improving the CS1 experience with pair programming. In *Proceedings of the 34th SIGCSE Technical Symposium on Computer Science Education (SIGCSE’03)*. ACM Press, 359–362.

NAPS, T. L. 2005. Jhavé – Addressing the need to support algorithm visualization with tools for active engagement. *IEEE Comput.- Graph. Appl. 25*, 5, 49–55.

NAPS, T. L. AND GRISSOM, S. 2002. The effective use of quicksort visualizations in the classroom. *J. Comput. Sci. Coll 18*, 1, 88–96.

NAPS, T. L., RÖSSLING, G., ALMSTRUM, V., DANN, W., FLEISCHER, R., HUNDHAUSEN, C., KORHONEN, A., MALMI, L., MCNALLY, M., RODGER, S., AND VELÁZQUEZ-ITURBIDE, J. Á. 2002. Exploring the role of visualization and engagement in computer science education. In *ITiCSE on Innovation and Technology in Computer Science Education (ITiCSE’02)*. (Working Groups Report). ACM Press, 131–152.

OECHSLE, R. AND MORTH, T. 2007. Peer review of animations developed by students. *Electron. Notes Theor. Comput. Sci. 178*, 181–186.

RAGONIS, N. AND BEN-ARI, M. 2005. On understanding the statics and dynamics of object-oriented programs. *SIGCSE Bull. 37*, 1, 226–230.

ROSCHELLE, J. 1996. Designing for cognitive communication: Epistemic fidelity or mediating collaborative inquiry. In *Computers, Communication & Mental Models*, D. L. Day and D. K. Kovacs Eds. Taylor & Francis, London, UK. 13–25.

RÖßLING, G. AND NAPS, T. L. 2002. A testbed for pedagogical requirements in algorithm visualizations. In *Proceedings of the Innovation and Technology in Computer Science Education (ITiCSE'02)*. ACM Press, 96–100.

SCAIFE, M. AND ROGERS, Y. 1996. External cognition: how do graphical representations work? *Int. J. Hum.-Comput. Stud. 45*, 2, 185–213.

SIMON, B., ANDERSON, R., HOYER, C., AND SU, J. 2004. Preliminary experiences with a tablet PC based system to support active learning in computer science courses. In *Proceedings of the 9th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education (ITiCSE'04)*. ACM, 213–217.

SPADA, H., MEIER, A., RUMMEL, N., AND HAUSER, S. 2005. A new method to assess the quality of collaborative process in CSCL. In *Computer Supported Collaborative Learning 2005: The Next 10 Years*, T. Koschmann, D. Suthers, and T. W. Chan Eds. Lawrence Erlbaum, Mahwah, NJ. 622–631.

SUTHERS, D. D. AND HUNDHAUSEN, C. D. 2003. An experimental study of the effects of representational guidance on collaborative learning processes. *J. Learn. Sciences 12*, 2, 183–219.

SUTHERS, D. D., HUNDHAUSEN, C. D., AND GIRARDEAU, L. E. 2003. Comparing the roles of representations in face-to-face and online computer supported collaborative learning. *Comput. Ed. 41*, 4, 335–351.

TEASLEY, S. 1997. Talking about reasoning: How important is the peer in peer collaboration. In *Discourse, Tools and Reasoning: Essays on Situated Cognition*, L. Resnick, R. Säljö, C. Pontecorvo, and B. Burge Eds. Springer, Berlin, Germany. 361–384.

WILLIAMS, L., KESSLER, R. R., CUNNINGHAM, W., AND JEFFRIES, R. 2000. Strengthening the case for pair programming. *IEEE Softw. 17*, 4, 19–25.

Received January 2008; revised September 2008; accepted November 2008