Algorithm Visualization in CS Education: Comparing Levels of Student Engagement

Scott Grissom
Grand Valley State University
grissom@gvsu.edu

Myles F. McNally
Alma College
mcnally@alma.edu

Tom Naps
University of Wis. - Oshkosh
naps@uwosh.edu

Abstract
Software technology for algorithm visualization (AV) has advanced faster than our understanding of how such technology impacts student learning. In this paper we present results of a multi-university study. We measured the effect of varying levels of student engagement with AV to learn simple sorting algorithms. These levels included: 1) not seeing any visualization, 2) simply viewing visualizations for a short period in the classroom, and 3) interacting directly with the visualizations for an extended period outside of the classroom. Our results show that learning increases as the level of student engagement increases. AV has a bigger impact on learning when students go beyond merely viewing a visualization and are required to engage in additional activities structured around the visualization. In particular, students who responded to questions integrated into the AV tool during their exploration of an algorithm showed the most improvement between a pretest and posttest.

Categories and Subject Descriptors
K.3.1 [**Computers and Education**]: Computer and Information Science Education

General Terms
Algorithms, Experimentation

Keywords
Algorithm visualization, sorting algorithms, computer science education

1. Introduction
Algorithm visualization (AV) depicts the execution of an algorithm as a discrete or continuous sequence of graphical images, the viewing of which is controlled by the user. Many algorithm visualization tools oriented toward computer science education have been developed and presented at recent SIGCSE and ITiCSE conferences. There have also been numerous empirical experiments attempting to prove the instructional effectiveness of AV. Hundhausen et al. [2002] presents a meta-study of twenty-one such experiments. Perhaps the most interesting observation growing out of this meta-study is that the determining factor in establishing the effective use of AV is not so much the features of the AV tool. Rather it is the manner and degree with which the learner becomes engaged with activities beyond merely watching the visualization. For example, in the twenty-one studies cited by Hundhausen et al., twelve involved such additional activities. Of those twelve, ten demonstrated improved learning at a statistically significant level. Of the nine studies in which learners merely watched different visual representations of the algorithm, only three showed a significant result.

Results such as these inspired the convening of a Working Group on "Improving the Educational Impact of Algorithm Visualization" at the June 2002 ITiCSE (Integrating Technology into Computer Science Education) conference in Aarhus, Denmark. The report of that group [Naps et al. 2003] identified an *engagement taxonomy* encompassing six different forms of learner engagement with visualization technology. The group's purpose in defining this taxonomy is to provide a framework for conducting empirical experiments that attempt to evaluate the instructional effectiveness of AV. The six categories in the working group's taxonomy are:

1. No viewing
2. Viewing
3. Responding
4. Changing
5. Constructing
6. Presenting

In this taxonomy, "no viewing" refers to instruction without any form of accompanying AV. “Viewing” can be considered the core form of AV engagement, since all other, more active, forms of engagement with visualization technology fundamentally extend some kind of viewing. Viewing by itself is the most passive of the forms of engagement. Aside from controlling a visualization’s execution and changing views, viewing by itself does not entail active involvement with a visualization. Hence, relative to Hundhausen's meta-study, "viewing" would be used to classify the nine experiments in which learners merely watched different visual representations of the algorithm being studied.

Category 3 in the engagement taxonomy is “Responding”. The key activity in this category is having the learner answer questions concerning the visualization while it is presented by the system. Answering these questions may be integrated into the AV system, or it may be in the form of questions administered via separate pencil-and-paper exercises that the learner completes while watching the algorithm execute in the AV system. Two of the studies cited in Hundhausen's meta-study used this form of engagement. The first of these was a study in which the researchers found that forcing students to do prediction during the animation of a depth-first search

ACM Symposium on Software Visualization, San Diego, CA

produced significantly better results on a post-test taken by students [Byrne et al. 1999]. Their methodology was to have students in a closed lab setting orally predict the behavior exhibited by the algorithm visualization. A second conflicting result was reported in a more recent study [Jarc et al. 2000]. This study used Jarc's Interactive Data Structure Visualizations (IDSV) software to automate what Byrne, Catrambone, and Stasko had students do orally. That is, the ISDV software system itself asked predictive questions of the students as they watched an algorithm. Hence students could (and were expected to) use the system on their own, outside of a closed lab setting. The students who used ISDV in the study did no better than students who were not using the system. Jarc, Feldman, and Heller hypothesize that this ineffectiveness is because poorer students merely treated the interactive questions as a game. Once they became lost in watching an algorithm, they completed the questions simply by making guesses.

Category 4 in the engagement taxonomy, “Changing,” entails modifying the visualization. The most common example of such modification is allowing the learner to change the input of the algorithm under study in order to explore the algorithm’s behavior in different cases. Work by Lawrence [1993; Lawrence et al. 1994] that was cited in Hundhausen's meta-study reports on two separate experiments in which learners who engaged in this changing mode of interaction performed significantly better than those who passively viewed an animation.

Category 5 in the engagement taxonomy is “Constructing”. In this form of engagement, learners construct their own visualizations of the algorithms under study. Although one would suspect that having students construct their own visualizations should greatly increase understanding, the evidence so far is unclear in this regard. One of the surprising non-significant results in Hundhausen's meta-study was inspired by Stasko's initially enthusiastic response to an experiment in which he had students construct their own visualizations using his Samba system [Stasko 1997]. Stasko reported anecdotal evidence that having students do this seemed to result in their learning the algorithm better than if they merely watched a visualization that was constructed for them. However, a more thorough follow-up study on having students “self-construct” a visualization versus having them actively view an expert-constructed visualization was conducted by [Hundhausen and Douglas 2000]. They found no significant difference. Hundhausen and Douglas attribute this result to the limited time frame (two-and-one-half hours) in which the two groups – the “self-constructors” and the “active-viewers” – had to work with the algorithm in question (a QuickSelect algorithm to find the kth minimum in an array).

Category 6 in the engagement taxonomy, “Presenting”, entails having students present a visualization to an audience for feedback and discussion. The visualizations to be presented may or may not have been created by the learners themselves. Apparently no empirical studies have yet been conducted that would indicate the effectiveness of this type of engagement.

The study we present in this paper uses the framework of the working group's taxonomy to compare the performance of three treatment groups having different levels of engagement with visualizations. These three levels are the "No viewing", "Viewing (only)", and "Responding" levels. Hence this study offers the potential of further illuminating the seemingly conflicting results reported in previous studies [Byrne et al. 1999; Jarc et al. 2000].

2. Materials And Method

2.1 Design of the Investigation

A between subjects design of three treatments was used. The treatments were identified based on the amount of interaction students had with visualizations of sorting algorithms. Students in treatment None (N=62) did not see any visualizations. Students in treatment Viewed (N=62) simply saw visualizations briefly in lecture that were controlled by the instructor. Students in treatment Responded (N=33) saw the visualizations in lecture but also interacted with the them during a closed lab.

Four faculty members from three universities taught seven sections. Two sections were for a CS 1 course and the remaining five sections were for a CS 2 course. Two professors taught two sections, one with Treatment None and one with Treatment Viewed. One professor taught two sections of Treatment Responded. One professor taught one section of Treatment None.

2.2 Materials

Students completed a pretest before reading any information about the sorting algorithms (Appendix A). Seven questions measured the student’s self-rating of visual learning preference. For example, “Assembling a bicycle from a diagram would be (easy or challenging)”. A response of ‘easy’ suggests the person has a preference for visual learning. Six additional questions resulted in a possible self-rating of 0 to 7. Four questions measured the student’s current knowledge about sorting algorithms. Each question related to the number of comparisons, swaps or passes for one of the algorithms under different input conditions. These questions can be answered with a general understanding of the algorithm. For example, “which algorithm(s) perform the same number of comparisons regardless of the starting condition?”

Instructors used identical lecture slides to describe each algorithm and provided a five-page study guide to each student. Students were encouraged to use only the study guide and not their textbook which was different for each university. Neither the lecture slides nor the study guide used visual representations of the algorithm.

The JHAVÉ visualization tool supplemented the lecture for Treatments Viewed and Responded [Naps et al. 2000]. Treatment Responded had direct access to JHAVÉ. They saw a text description of the algorithm, an interactive visualization, and an integrated on-line quiz (Figure 1).

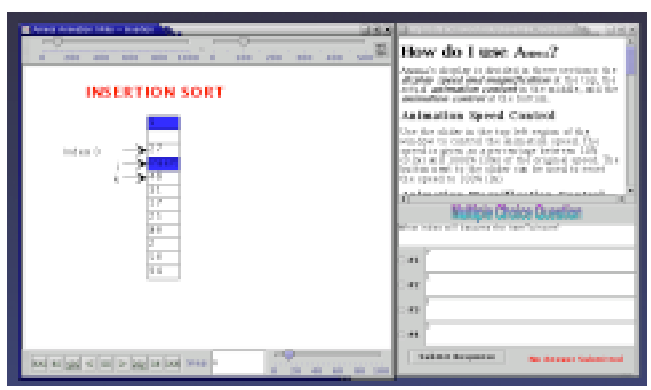

Figure 1. Typical JHAVÉ screen shot.

A posttest included twelve questions about the sorting algorithms and a brief survey about their experience with the learning tool. Four of the questions were identical to the sorting questions on the pretest. Four questions required students to write or read code. We label these Code Questions versus Visual Questions. Visual Questions can be answered with a mental model of an algorithm that does not rely on the underlying implementation.

2.3 Procedure

Students completed the pretest before covering any of the material: bubble sort, selection sort and insertion sort. Each professor used the same lecture slides to describe the algorithms and students were given a printed study guide to use instead of their textbook.

Students attended a lecture for one class session that covered all three sorting algorithms. The posttest was administered one week later.

Instructors used a common grading metric to insure the exams were graded consistently. Several of the questions were multiple choice so there was not ambiguity on how to grade these questions. Full credit for the correct answer and no credit otherwise. Other questions had specific answers that were either correct for full credit, incorrect for no credit or partially correct for an agreed amount of credit.

3. Results

We used the non-parametric Kruskal-Wallis and Mann-Whitney tests for comparisons between treatments unless otherwise noted. A t-test was not used since the assumption of normality was not met for most data. A 5% significant level was used. Correlation was measured with Spearman's Rho.

Figure 2 shows a significant difference between treatments with respect to pretest scores ($p=0.001$). Treatment None ($N=51$) had a mean score of 10.6. Treatment Viewed ($N=51$) had a mean score of 7.3. Treatment Responded ($N=33$) had a mean score of 5.1. Post hoc tests indicated the only pair-wise significant difference was between None and Responded ($p \leq 0.0001$).

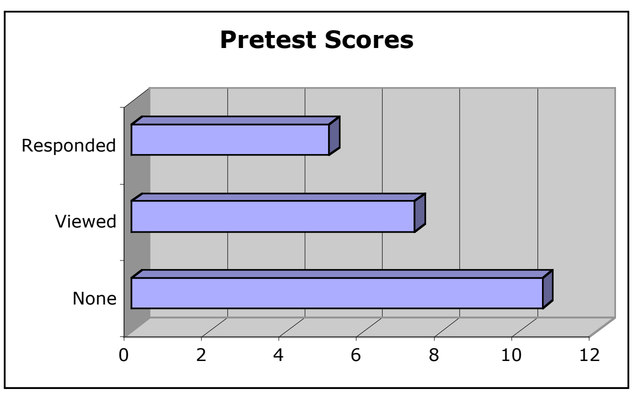

Figure 2. Mean scores of the pretest. Maximum = 33.

Assumptions of normality were met for the ACT college prep test scores and an ANOVA revealed no significant difference between treatments. Treatments None and Viewed each had mean ACT scores of 25.4. Mean ACT score of Responded was 23.6.

Four of the posttest questions were identified as coding questions. There is little reason to suspect that these questions would be improved by seeing visualizations since the visualizations did not show corresponding code as the algorithm performed. There was no significant difference between treatments on coding questions ($p = 0.194$). Treatment None ($N = 62$) had a mean score of 25.5. Treatment Viewed ($N = 62$) had a mean score of 27.8. Treatment Responded ($N = 33$) had a mean score of 28.5.

Eight posttest questions were identified as visual questions. The four pretest questions were a subset of the visual questions. The pretest score was subtracted from the posttest visual questions to measure improvement (Figure 3). There was a significant difference between treatments ($p < 0.005$) using ANOVA. Treatment None ($N = 51$) had a mean score of 35.9. Treatment Viewed ($N = 51$) had a mean score of 39.8. Treatment Responded ($N = 33$) had a mean score of 46.3. Post hoc tests revealed that the only significant pair-wise difference was between Responded and None ($p < 0.005$).

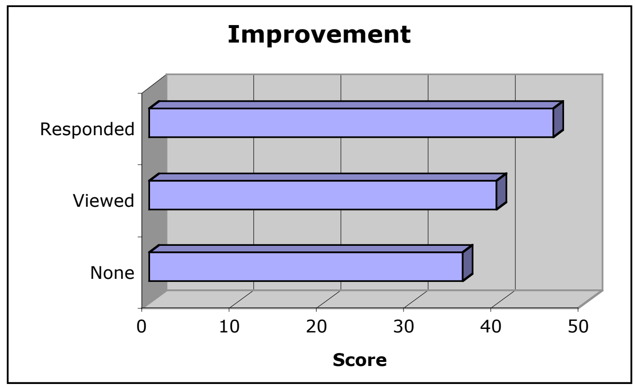

Figure 3. Improved performance on visual questions between the pretest and posttest.

Students completed a self-rating of preferred learning style on the pretest. Ratings ranged from 0 for no preference to visual learning to a maximum of 7. There was no correlation between visual learning style and performance on the visual questions.

Students in the Viewed and Responded treatments evaluated the effectiveness of the visualization on a scale from one (not effective) to five (very effective). There was no significant difference between Viewed (3.7) and Responded (3.8).

4. Discussion

4.1 What were our significant results?

Our data show that learning improves as the level of student engagement with AV increases. Learning was measured by subtracting pretest from posttest scores. Improvement occurred between each level of engagement (Figure 3). The improvement between not viewing a visualization and interacting with one was statistically significant.

4.2 Did the treatment groups have similar academic abilities?

Yes. It is important to determine that each treatment has participants with similar academic skills and experience. This is especially important when the participants are at three different schools. We used the ACT college entrance exam as a measure of general scholastic ability. There was no significant difference between treatments. All students were CS majors taking their first or second course in the programming sequence. Results from the pretest indicated that Treatment None had more initial knowledge about the algorithms. This advantage was factored out by measuring improvement between the pretest and posttest instead of only using the raw posttest scores.

It is also interesting to note that both Treatment None and Treatment Viewing were composed of students in a CS2 course while Treatment Responding was composed of students in a CS1 course. Therefore, Treatment Responded may have had slightly less academic experience than the other groups. If the situation had been reversed and the None group had less experience than the Responded group, one could claim that these covariant factors may have affected the results. That the Responding group, despite less experience, went on to outperform the None group lends further weight to the significance of our results.

4.3 Was there a difference in learning with respect to different types of questions?

Yes. The four coding questions relied on traditional programming skills. One might suspect that such skills would not be improved with a visual representation of the algorithms. Indeed, in our study, there was no difference between treatment groups on coding questions. This confirms that there is some learning that can be improved with visualizations and some that might not.

The issue of trying to determine the type of learning that might be improved with AV is also addressed in the ITiCSE 2002 working group report [Naps et al. 2003]. This report suggests a metric for different types of learner understanding in a study such as ours. This metric is the well-known Bloom Taxonomy [1]. Bloom's taxonomy structures a learner's depth of understanding along a linear progression of six increasingly sophisticated levels:

Level 1: The *knowledge level*. It is characterized by mere factual recall with no real understanding of the deeper meaning behind these facts.

Level 2: The *comprehension level*. At this level, the learner is able to discern the meaning behind the facts.

Level 3: The *application level*. Now the learner can apply the learned material in specifically described new situations.

Level 4: The *analysis level*. The learner can identify the components of a complex problem and break the problem down into smaller parts.

Level 5: The *synthesis level*. The learner is able to generalize and draw new conclusions from the facts learned at prior levels.

Level 6: The *evaluation level*. The learner is able to compare and discriminate among different ideas and methods. By assessing the value of these ideas and methods, the learner is able to make choices based on reasoned arguments.

With respect to the understanding of introductory sorting algorithms measured in our study, we would claim that only the first three levels of Bloom's taxonomy are relevant. That is, our expectations were that the student should understand the "recipe" behind each of the sorting algorithms (Level 1), how this recipe results in data swaps and comparisons being made for particular data sets (Level 2), and finally how to transfer understanding of the recipe into specific decisions about how to code the sorting algorithm (Level 3). Under this interpretation, our results indicate that visualizations help learning at Levels 1 and 2, but not at Level 3. What does this suggest about teaching strategies that incorporate visualizations? Visualizations seem to help with those types of understanding that do not require deep and reflective thought. Hence an instructor might be advised to use visualizations as a means to insure that students understand the "basics" without having to waste substantial class time on such issues. Consequently instructor-student interaction during regular class time can be devoted to higher levels on the Bloom scale.

4.4 What affect did student motivation have on performance?

Data from one of the universities is not reported in Section 3 because we were concerned the data may have been inadvertently compromised. These students were assigned to the None Treatment. The instructor explained to the small number (N=9) about the different treatment groups at other Universities. They took the study as a personal challenge to outperform the other groups. We were concerned that this extra incentive, of which the other treatment groups were not aware, may have affected their performance. This was indeed the case. Performance of this group (Treatment None X) was higher than any other group (Figure 4). However, even when these data are included in the full set, we still see the same significant results. Despite this, we chose not to include the data in the reported set. Another factor determining this choice was that we did not have ACT scores for students in Treatment None X.

The very fact that this outlier group performed well under the guise of a "challenge" from their instructor is indicative of what is at the same time both a flaw with our study and an interesting covariant factor. We specifically chose not to control the "time on task" that the groups spent trying to learn the material. To what extent is the statistically better performance of the Responding Treatment (as well as the small outlier group) a consequence of their spending more time on task rather than their interacting with the visualizations? The data we collected does not allow us to reliably answer this question. In fact, our previous experience suggests that self-reported data for study times is not a reliable indication of performance [Naps and Grissom 2002].

4.5 How do our results compare to previous studies?

In Section 1 we described conflicting results regarding the "Responding" level of engagement in previous studies [Byrne et al. 1999; Jarc et al. 2000]. The former found that learners achieved significantly better results in their understanding of the algorithm while the latter found no significant difference. Our results reinforce the findings of the Byrne study.

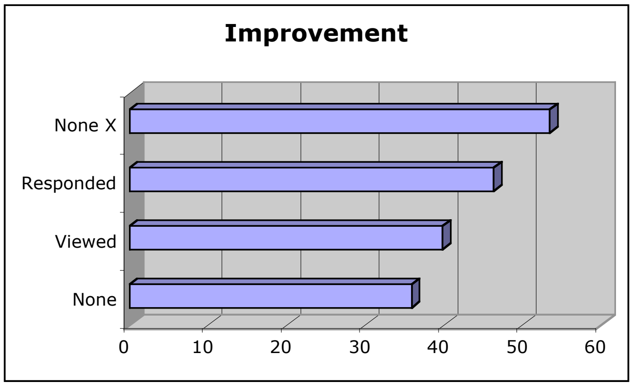

Figure 4. Improved performance on Visual Questions between the pretest and posttest with the excluded data set.

While all three studies used the Responding level of engagement, there is an interesting spectrum in how the Responding level was administered. The Byrne study was conducted in a very controlled fashion in which the participating learners were under careful observation during their limited time with the AV system. In the Jarc study, students were given the opportunity to use the system on their own throughout an entire course. Our study falls somewhere in between these studies in terms of the degree of control we exercised over students. Our Responding Treatment group used the AV tool in a closed lab setting. However, this only represented the learners' initial encounter with the AV tool. We also gave students additional time to use the tool on their own. This resembled the Jarc study, except that we limited our period of observation to one topic covered in one week instead of multiple topics covered over a semester-long time frame. On average, our students reported using the AV tool for thirty one minutes on their own.

These studies illustrate a "spectrum of control" exercised by the instructors over the learners. Perhaps not surprisingly, the two studies in which more control was exercised resulted in positive learning results. This is an indication that, if we expect relatively naive students to benefit from AV, we must carefully guide students in their explorations with the AV tool.

4.6 What affect did preferred learning style have on response to the visualizations?

Researchers (for example, [Felder 1993]) have suggested that visualizations are more helpful to visual learners than non-visual learners. Our data does not support this suggestion. There was no correlation between performance on the visual questions and the preferred learning style. Our intuition is that there is a correlation between visual learners and their response to AV. Perhaps a different learning style inventory would provide different results. We plan to compeare the inventories used in previous studies to the one we used in this study. It would be useful to identify a common learning style inventory for researchers in this field to use.

4.7 What are the challenges of collecting data in the classroom?

We continue to learn that conducting controlled studies in the classroom is challenging [Naps and Grissom 2002]. The obvious advantages to this approach are the large number of participants and experience with a realistic learning environment. The disadvantage is that the experimental variables are not as controlled as they could be in an artificial environment. We believe results from both types of studies contribute to our understanding.

Moreover, there are "human subjects" issues that emerge in a study such as ours when it is linked to students' grades in an actual course. In particular, if we as instructors suspect that more active levels of engagement with visualizations help students learn, then how do we justify withholding one group's access to such a learning resource? All three of the authors had to carefully approach their institutional human subjects committees on this issue.

5. Conclusions

The true value of using visualizations may lie not in their content but rather in their serving as a motivational factor to make students work harder. A recent article by Young [2002] laments the decreasing amount of time that today's college students spend studying outside of class. Young posits that "students today are so accustomed to distraction -- and bombardment with media images -- that they find it harder to concentrate than students in the past." If Young is correct and if visualizations serve to better focus the attention of the modern student, then that may be their greatest value.

To the degree that visualizations helped our students learn material, it must be emphasized that we developed substantial materials to accompany the visualizations. These included lecture notes and a printed study guide in which we were careful to present versions of the sorting algorithms that were consistent with what the students would see in the visualizations. If students interact with visualizations that are not consistent with algorithms they see in their textbooks, then the visualizations may actually serve to confuse more than help them. An unavoidable consequence of this is that "educators who are engaged in visualization need to devote more time to the (development of) accompanying instructional materials, rather than having a single-minded focus on the graphics of visualization" [Naps and Grissom 2002].

6. Acknowledgements

This work was supported by the National Science Foundation, NSF DUE-0126494. We appreciate Dr. Phyllis Curtiss and Crystl Oberth at the GVSU Statistical Consulting Center for their guidance with data analysis and interpretation of the results. We also appreciate the effort of our students and their willingness to use JHAVÉ .

7. References

BLOOM, B.S. (Ed.) 1956. Taxonomy of educational objectives: The classification of educational goals: Handbook I, cognitive domain. New York; Toronto: Longmans, Green.

BYRNE, M. D., CATRAMBONE, R. and STASKO, J. T. 1999. "Evaluating Animations as Student Aids in Learning Computer Algorithms", *Computers & Education*, 33, pp. 253-278.

FELDER, R. M. 1993. "Reaching the Second Tier — Learning and Teaching Styles in College Science Education," *Journal of College Science Teaching*, pp 286-290, 23 (5).

HUNDHASUEN, C. D. and DOUGLAS, S. A. 2000. "Using Visualizations to learn algorithms: Should students construct their own, or view an expert's?", In *2000 IEEE Symposium on Visual Languages* (pp. 21-28). Los Alamitos, CA: IEEE Computer Society Press.

HUNDHAUSEN, C. D., DOUGLAS, S. A., and STASKO, J. T. 2002. "A Meta-Study of Algorithm Visualization Effectiveness," *Journal of Visual Languages and Computing,* 13(3), pp. 259-290.

JARC, D. J., FELDMAN, M.B. and HELLER, R. S. 2000. "Accessing the Benefits of Interactive Prediction Using Web-based Algorithm Animation Courseware," in *Proceedings of the SIGCSE Session*, ACM Meetings, Austin, Texas.

LAWRENCE, A. W. 1993. "Empirical Studies of the value of algorithm animation in algorithm understanding," unpublished Ph.D. dissertation, Department of Computer Science, Georgia Institute of Technology.

LAWRENCE, A.W., BADRE, A, and STASKO, J. 1994. "Empirically Evaluating the use of Animations to Teach Algorithms," in *Proceedings of the 1994 IEEE Symposium on Visual Languages*, St. Louis, MO, pp. 48-54.

NAPS, T. and GRISSOM, S. 2002. "The Effective Use of Quicksort Visualizations in the Classroom," *Journal of Computing Sciences in Colleges*, CCSC, pp. 88-96.

NAPS, T (co-chair), RÖßLING, G. (co-chair), ALMSTRUM, V., DANN, W., FLEISCHER, R., HUNDHAUSEN, C., Korhonen, A., MALMI, L., McNALLY, M., RODGER, S., VELÁZQUEZ-ITURBIDE, J.A. 2003. "Exploring the Role of Visualization and Engagement in Computer Science Education," Report of the Working Group on "Improving the Educational Impact of Algorithm Visualization", to appear in *ACM SIGCSE Bulletin*, June.

NAPS, T., EAGAN, J, and NORTON, L. 2000. "JHAVÉ -- An Environment to Actively Engage Students in Web-based Algorithm Visualizations", in *Proceedings of the SIGCSE Session*, ACM Meetings, Austin, Texas.

STASKO, J. 1997. "Using Student-Built Algorithm Animations as Learning Aids" in *Proceedings of the SIGCSE Session,* ACM Meetings, San Jose, CA.

YOUNG, J. R.. 2002. "Homework? What Homework?", *Chronicle of Higher Education*, December 6.

Appendix A
Sorting Algorithms Pre-Test

Learning Style

Completing the following questions will help us identify your preferred learning style. There are no right or wrong answers. Check ONLY ONE response for each question.

1. To solve a math problem, I would prefer to
____ draw or visualize the problem.
____ study a sample problem and use it as a model.

2. To remember things best, I
____ create a mental picture.
____ write it down.

3. Assembling a bicycle from a diagram would be
____ easy.
____ challenging.

4. I prefer classes in which I
____ handle equipment or work with models.
____ participate in a class discussion.

5. To understand and remember how a machine works, I would
____ use a diagram.
____ write notes.

6. I prefer
____ drawing or working with my hands.
____ speaking, writing, and listening.

7. If you were trying to locate an office on an unfamiliar campus, would you prefer someone to
____ draw you a map.
____ tell you how to find the office.

Sorting Algorithms

This portion of the quiz is designed to measure any previous experience you have with sorting algorithms. You probably will not know the correct answers but do you best. When reading the following questions, be aware of the distinctions between array **passes,** entry **comparisons** and entry **swaps**.

1. Apply the **bubble sort algorithm** to the following array. Use the empty table cells to show the contents after each pass. The array may or may not be sorted after the third pass.

Array Index	0	1	2	3	4
Initial State of Array	98	78	53	32	17
show the array contents **after the entire first pass**					
show the array contents **after the second pass**					
show the array contents **after the third pass**					

2. Consider the **improved bubble sort** on the following array. How many **passes** will the algorithm make before stopping?

Number of passes _______

Array Index	0	1	2	3	4
Initial State of Array	89	15	53	72	85

3. Consider Selection Sort, Insertion Sort, Bubble Sort and Advanced Bubble Sort. Which of the four sorting algorithms (if any) will always perform the same number of **passes** and **comparisons** regardless of the initial array? The correct answer might include more than one algorithm.

4. Given an initial array with N items that is already sorted, how many **comparisons** (not passes) would **insertion sort** make?

Number of Comparisons ______