Aalto University

OPEN ACCESS

This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

McCartney, Robert; Moström, Jan Erik; Sanders, Kate; Seppälä, Otto

# Questions, annotations and institutions: observations from a study of novice programmers

*Published in:*
Kolin Kolistelut-Koli Calling 2004, Fourth Finnish/Baltic Sea Conference on Computer Science Education, Koli, Finland, October 1-3, 2004

Published: 01/01/2004

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

*Please cite the original version:*
McCartney, R., Moström, J. E., Sanders, K., & Seppälä, O. (2004). Questions, annotations and institutions: observations from a study of novice programmers. In A. Korhonen, & L. Malmi (Eds.), *Kolin Kolistelut-Koli Calling 2004, Fourth Finnish/Baltic Sea Conference on Computer Science Education, Koli, Finland, October 1-3, 2004* (pp. 11-19). Helsinki University of Technology.

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

# Questions, Annotations, and Institutions: observations from a study of novice programmers

Robert McCartney
*Dept. of Computer Science and Engineering, Univ. of Connecticut, Storrs, CT, USA*

Jan Erik Moström
*Department of Computing Science, Umeå University, Umeå, Sweden*

Kate Sanders
*Dept. of Math and Computer Science, Rhode Island College, Providence, RI, USA*

Otto Seppälä
*Dept. of Computer Science and Engineering, Helsinki Univ. of Technology, TKK, Finland*

robert@cse.uconn.edu

**Abstract**

This paper examines results from a multiple-choice test given to novice programmers at twelve institutions, with specific focus on annotations made by students on their tests. We found that the question type affected both student performance and student annotations. Classifying student answers by question type, annotation type (tracing, elimination, other, or none), and institution, we found that tracing was most effective for one type of question and elimination for the other, but overall, any annotation was better than none.

## 1 Introduction

In summer, 2004, a working group at the ITiCSE conference in Leeds, UK, examined the code reading and understanding of novice computer programmers. The analysis was based on data collected from twelve institutions in Australia, Denmark, England, Finland, New Zealand, Sweden, the United States, and Wales. These data were based on a multiple-choice test administered to beginning programming students, and included the student answers for all of the questions. For a subset of the students, interview transcripts and the actual test forms with any annotations used during the test were also collected. The working group analyzed a broad range of issues: the kinds of questions, the performance of students by institution and quartile, the sorts of annotations used and their general effectiveness, student test-taking strategies observed, and others(Lister et al., 2004).

In this paper, we look in detail at a small slice of these issues: the interrelationships observed between the kinds of annotations used by the students, the style and difficulty of the individual questions, and the institutions where the students were tested. In particular, we would like to determine how the likelihood of answering a question correctly is affected by the various kinds of annotations used.

## 2 Classifying annotations

An annotation, referred to as a "doodle" by the Working Group, was defined to be any kind of marking by a student on his or her exam paper. In Figure 1 we can see an example where the user made a number of marks: numbers written over variables, numbers written under array elements, and many assignment statements setting variables to constants. In this case, the assignments show the student keeping track of variables as the loop is executed–a graphical record of the tracing process.

Two of the Working Group members did a data-driven classification of the doodles. This classification was later independently verified by three other members (see (Lister et al., 2004) for details). The classification is given in Table 1. Using this, the researchers classified 56 complete exams: the three from each institution corresponding to the students who were

interviewed about their tests, plus 20 others chosen at random from the six researchers who had other tests with them in Leeds.

**Table 1**: Categorization of annotations. % is percentage of questions showing each type.

| Name | Code | Description | % |
|---|---|---|---|
| Blank page | B | No annotations for this question | 38 |
| Synchronized trace | S | Shows values of multiple variables changing, generally in a table | 11 |
| Trace | T | Shows values of a variable as it changes (more than 1 value) or a variable's value is overwritten with new value | 32 |
| Odd Trace | O | Something that appears to be a trace but neither S or T, such as linking representations with arrows | 3 |
| Alternate answer | A | Student changed their answer to the question | 4 |
| Ruled out | X | One or more alternative answers crossed out so it answer appeared to be selected by elimination | 9 |
| Computation | C | An arithmetic or boolean computation (not rewrite of comparison) | 4 |
| Keeping tally | K | Some value counted multiple times, variable not identified | 1 |
| Number | N | Shows single variable value, most often in comparison | 28 |
| Position | P | Picture of correspondence between array indices and values | 11 |
| Underlined | U | Part of question underlined for emphasis | 7 |
| Extraneous marks | E | Markings that appear meaningless or ambiguous (could not be characterized). Includes arrows, dots, and so forth | 13 |

As the example in Figure 1 shows, an answer can contain several different doodle categories: N, P, and T for this question.

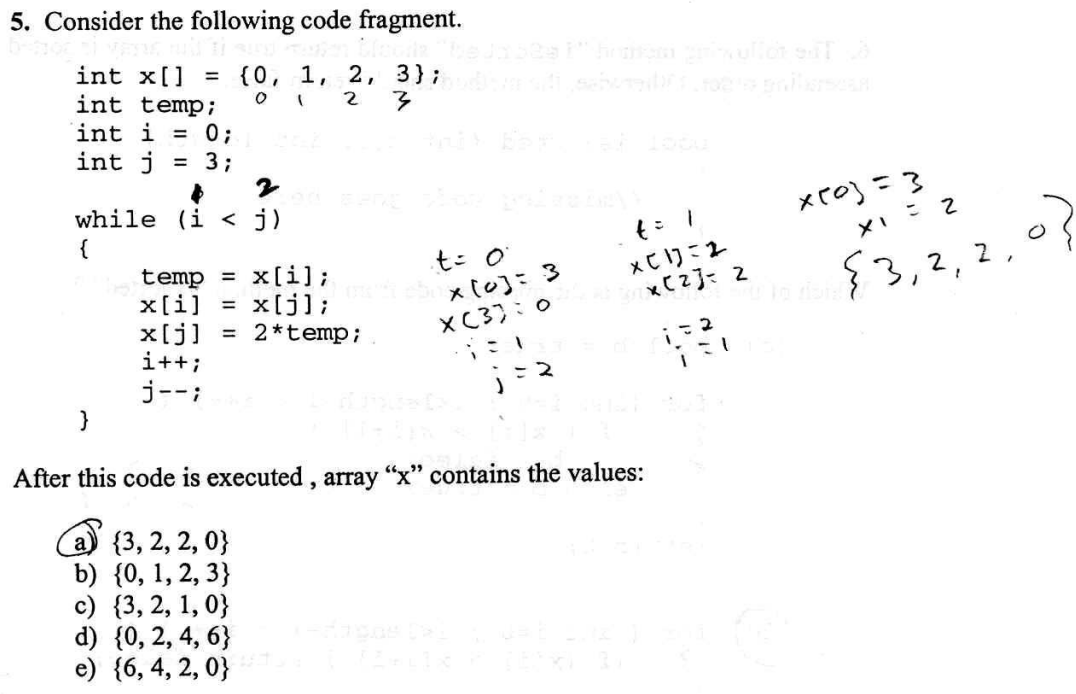



**Figure 1**: Example of doodles on test from question 5, showing annotations N, P, and T.

## 3 Classifying the exam questions

The multiple-choice exam used for this project consisted of twelve questions involving a variety of array-processing tasks, such as comparing two arrays in order to find elements that are contained in both, testing to determine whether an array is sorted, filtering some of the elements of one array into another, searching for a given element in an array, and deleting the element in a given position from an array.

Each question on the test can be classified into one of two categories, *fixed-code* questions, where the student is given a code fragment and asked questions about the result of executing it, and *skeleton-code* questions, where the student is given an incomplete code fragment, and asked to complete it so it will perform a given task. There were seven fixed-code and five skeleton-code questions on the exam.

Questions 3 and 5 in Figure 2 are representative fixed-code questions. They give a single piece of code and ask the student about what is true after the code is executed. The style of all the fixed-code questions is the same as these: "Consider the following code fragment: *[code fragment]* What is the value of *[some int or array variable]* after this code is executed?" And each of the possible answers is a constant, either the value of an integer variable (as in question 3) or a list of the values in an array (as in question 5). Other than code, they require very little reading.

Question 8, also shown in Figure 2, is a representative skeleton-code question. Like Question 8, the skeleton-code questions generally provide a description of what the code is to accomplish, a code fragment containing one or more blanks, and a set of choices to fill in the blanks. The choices are all code fragments themselves, as opposed to the numeric values used in the answers to the fixed-code questions.

The students' performance was noticeably different on the two types of questions. Overall, our students had an average score of 61%, indicating that they do not have a strong grasp of the basic knowledge in these areas. Students did better on the fixed-code questions than on the skeleton-code questions, however. On the fixed-code questions, from 61% to 73% of all the answers were correct (depending on the question). On the skeleton code questions, the percentages were 35%, 43%, 46%, 58%, and 72%–all but one of them worse than any of the fixed-code questions.

## 4 Analysis

Two issues complicated our analysis of these data. First, the large majority of the annotations were done on the fixed-code questions, as can be seen in Table 2. The difference between the fixed-code and skeleton-code questions is striking: 77% of the fixed-code questions are annotated, as opposed to 41% of the skeleton-code questions. Indeed, if we were to group the questions on the basis of how often they are annotated, without looking at the questions themselves, we would have the same groups: Questions 1-5, 7, and 10, the fixed-code questions (68%-88% annotated) and Questions 6, 8-9, and 11-12, the skeleton-code questions (38%-45% annotated). Because the two groups were annotated so differently, there was the danger

**Table 2**: Number and and percentage of annotated questions observed. These are based on analysis of 672 questions (56 of each question).

| | Question | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| count | 38 | 44 | 40 | 48 | 49 | 23 | 41 | 25 | 22 | 40 | 25 | 21 |
| percent | 68% | 79% | 71% | 86% | 88% | 41% | 73% | 45% | 39% | 71% | 45% | 38% |
| Q type | FC | FC | FC | FC | FC | SC | FC | SC | SC | FC | SC | SC |

that overall conclusions would be determined by the data from the annotations of the fixed-

**Question 3.** Consider the following code fragment:

```
int [] x = {1, 2, 3, 3, 3};
boolean b[] = new boolean [x.length];
for ( int i = 0; i < b.length; ++i )
    b[i] = false;
for ( int i = 0; i < x.length; ++i )
    b[ x[i] ] = true;

int count = 0;

for (int i = 0; i < b.length; ++i )
    if ( b[i] == true ) ++count;
```

After this code is executed , "count" contains:

a) 1
b) 2
c) 3
d) 4
e) 5

**Question 5.** Consider the following code fragment:

```
int [] x = {0, 1, 2, 3};
int temp;
int i = 0;
int j = x.length-1;
while (i < j) {
    temp = x[i];
    x[i] = x[j];
    x[j] = 2*temp;
    i++;
    j--;
    }
```

After this code is executed, array "x" contains the values:

a) {3, 2, 2, 0}
b) {0, 1, 2, 3}
c) {3, 2, 1, 0}
d) {0, 2, 4, 6}
e) {6, 4, 2, 0}

**Question 8.** If any two numbers in an array of integers, not necessarily consecutive numbers in the array, are out of order (i.e. the number that occurs first in the array is larger than the number that occurs second), then that is called an inversion. For example, consider an array "x" that contains the following six numbers:

```
4   5   6   2   1   3
```

There are 10 inversions in that array, as:

```
x[0]=4   >   x[3]=2
x[0]=4   >   x[4]=1
x[0]=4   >   x[5]=3
x[1]=5   >   x[3]=2
x[1]=5   >   x[4]=1
x[1]=5   >   x[5]=3
x[2]=6   >   x[3]=2
x[2]=6   >   x[4]=1
x[2]=6   >   x[5]=3
x[3]=2   >   x[4]=1
```

The skeleton code below is intended to count the number of inversions in an array "x":

```
int inversionCount = 0;
for ( int i=0 ; i<x.length-1 ; i++ )
{
   for  ( xxxxxx )
   {
      if (  x[i] > x[j] ) ++inversionCount;
   }
}
```

When the above code finishes, the variable "inversionCount" is intended to contain the number of inversions in array "x". Therefore, the "xxxxxx" in the above code should be replaced by:

```
a)   for ( int j=0   ; j<x.length   ; j++ )
b)   for ( int j=0   ; j<x.length-1 ; j++ )
c)   for ( int j=i+1 ; j<x.length   ; j++ )
d)   for ( int j=i+1 ; j<x.length-1 ; j++ )
```

**Figure 2**: Questions 3, 5, and 8.

code questions. Accordingly, we considered the fixed-code and skeleton-code questions both together and as two separate groups.

The second issue that complicates this analysis is that the classifications are not disjoint; with the exception of Blank, any combination of annotations can occur on any given question. The observed non-blank questions had from 1 to 5 different annotation types represented, with an average of approximately 2. As an example of this problem, consider class N, numbering, which was relatively common (appearing in just over 28% of the questions). 90% of the time that there was numbering, however, there were other classes as well; 77% of the time at least one tracing type was also present.

To resolve this issue, we created four disjoint categories, reclassifying each question as

Blank, Some Tracing (S,T, and O, but not A or X), Elimination (A or X), or Other (everything else). The rationale for these categories is that tracing and process of elimination are recognizable strategies, and, with Blank, cover 89% of the observations.

We then counted the number and percentage of time the answers were correct for each category, for fixed-code questions, skeleton-code questions, and for all questions taken together. The results are given in the following tables.

**Table 3**: Percentages of annotation types for each question type. Numbers in parentheses present the counts (questions having this annotation, all questions).

| Question type | Blank | Some tracing | Elimination | Other |
|---|---|---|---|---|
| fixed-code | 23 (92 of 392) | 61 (238 of 392) | 6 (25 of 392) | 9 (37 of 392) |
| skeleton-code | 59 (164 of 280) | 6 (17 of 280) | 21 (59 of 280) | 14 (40 of 280) |
| all | 38 (256 of 672) | 38 (255 of 672) | 13 (84 of 672) | 11 (77 of 672) |

**Table 4**: Percentages correct, by question and annotation type. Numbers in parentheses present the counts (correct and total answers).

| Type | Blank | Some tracing | Elimination | Other | Total |
|---|---|---|---|---|---|
| fixed | 50 (46 of 92) | 76 (180 of 238) | 48 (12 of 25) | 54 (20 of 37) | 66 (258 of 392) |
| skeleton | 50 (82 of 164) | 65 (11 of 17) | 61 (36 of 59) | 55 (22 of 40) | 54 (151 of 280) |
| all | 50 (128 of 256) | 75 (191 of 255) | 57 (48 of 84) | 55 (42 of 77) | 61 (409 of 672) |

These tables illustrate a number of things:

1. Overall, answers showing explicit tracing are the most likely to be correct: 75% are correct, compared with 50% for questions without annotation. We see similar results for FC and SC questions handled separately.

2. Overall, elimination is the second-most effective strategy.

3. Across the board, both overall and for the fixed-code and skeleton-code subsets of the data, any form of annotation, even Other, is better than no annotation at all.

4. The frequency of tracing is much lower for skeleton code questions than fixed code questions.Although tracing is used less, it is still highly effective for the skeleton-code questions when it is used.

5. The frequency of elimination is much higher for the skeleton code questions (where it is the most common annotation) than for the fixed code questions. Its effectiveness for skeleton code questions is slightly better than tracing, and much better than no annotations.

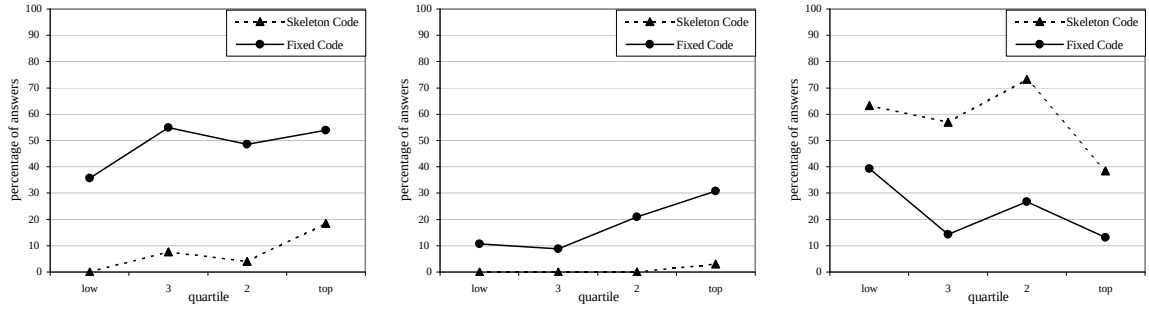
6. Skeleton-code questions are much more likely than fixed-code questions not to be annotated at all.

## 5 Annotations and overall performance

In the working group paper (Lister et al., 2004), the authors were able to extract differences between questions by observing what students in different quartiles (based on test score) chose in each question as their answer. We applied similar techniques here, with the expectation that students who do well on the exam would be more likely to use the effective annotation techniques. Results from three annotation types can be found in figure 3.

We found that tracing individual variables (T) was used about the same for the top three quartiles for fixed-code questions, but for skeleton-code questions, the top quartile was twice

as likely to trace as the second and third. The increased likelihood for the top quartile to use synchronized traces (S) relative to the lower quartiles was even greater, although the overall use was lower than tracing for the top quartile–possibly because not all questions required multiple variables to be traced. The data related to blank questions are less easily explained, however, as the frequency of blank questions does not increase monotonically as we go from the top to the bottom quartile – the second quartile students have relatively more blanks than the third quartile for fixed-code questions, and more than the third or fourth for skeleton-code questions.




(a) Tracing a single variable (b) Synchronized tracing (c) Blank

**Figure 3**: Percentage of an annotation type used in different quartiles

## 6 Annotations and institutions

While the popularity of annotations varied substantially from institution to institution, they led to higher scores in almost every institution. Table 5 shows the frequency of annotations for the 12 institutions (identified by letter due to confidentiality requirements).

**Table 5**: Percentage of questions with any annotations, by institution and question type.

| Question type | Institution | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | E | J | L | S | Q | H | O | A | T | N | C | P |
| all | 28 | 36 | 50 | 51 | 58 | 60 | 69 | 70 | 87 | 89 | 89 | 92 |
| fixed code | 36 | 55 | 62 | 71 | 71 | 89 | 81 | 89 | 98 | 86 | 100 | 100 |
| skeleton code | 17 | 10 | 33 | 23 | 40 | 20 | 53 | 43 | 73 | 93 | 73 | 80 |

These are large differences. Overall, the percentage of questions with annotations varies from 28 to 92. The percentages of fixed-code questions with annotations range from 36 to 100, and for skeleton-code questions, from 10 to 93.

To try to isolate the performance effects of annotation, we examined the performance difference between "strategic" annotations (tracing or elimination) and no annotations for each institution. (Tracing and eliminations were pooled since the numbers of observations at each institution are rather low). These data are given in Table 6. This table further supports the inference that students who annotate their exams tend to perform better: four of the five highest averages are from institutions in the top five in annotation frequency. In addition, in ten of the twelve institutions, annotated questions were more often correct than "blank" questions, which would indicate that the positive effect of tracing is fairly universal. (It should be noted that the number of blank questions is extremely low for institutions T,N,C, and P (9, 4, 4, and 3 respectively), so the comparisons for those institutions are of dubious value.)

**Table 6**: Percentage of questions correctly answered by institution, for Some tracing or elimination and blank. Average score is based on all questions at the institution.

| | Institution | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | E | J | L | S | Q | H | O | A | T | N | C | P |
| Some tracing or elimination | 87 | 77 | 87 | 54 | 72 | 59 | 63 | 76 | 78 | 63 | 76 | 58 |
| Blank | 62 | 48 | 33 | 66 | 53 | 21 | 45 | 38 | 67 | 100 | 50 | 0 |
| Average score | 68 | 57 | 58 | 60 | 61 | 42 | 56 | 63 | 78 | 67 | 75 | 39 |

## 7 Discussion: making sense of these results

Some of the results make obvious sense: for example, tracing through the code on paper helps; the proportion of correct answers where there is tracing is much higher than where there is not. One result seem counter intuitive at first glance: students use tracing less on the harder questions, where such strategies might be expected to be effective. We can offer three possible explanations:

**Too much work** Answering a skeleton-code question by tracing through the different alternatives would mean four or five times the work compared to a fixed-code question. Many students seem to realize this and instead chose to change their problem solving strategy: they begin by reading the question and forming a hypothesis of how the program should work, then look for boundary values that can be used to rule out different alternatives, and finally they check whether the selected alternative seem to work. This strategy would also explain the increase in eliminations for skeleton-code questions as shown in Table 3.

**Too abstract** The fixed-code questions are simple in the respect that they only require a knowledge of the syntax/semantics of the language and the ability to carefully trace the values of different variables. The skeleton-code questions are different in that they give a textual description of the problem, which the students have to translate into some problem understanding, they then have to form a hypothesis of a possible solution based on the alternatives, and then evaluate the different alternatives to find the correct one possible by creating one or more test cases (compare this to the software comprehension model described in (Boehm-Davis, 1988)).

**Lack of representation** Closely related to the explanation above is the lack of representation. As described above the student is required to *understand* the problem description for skeleton-code questions. This might be difficult to do without representing the task at the abstract level, a skill that most novices apparently do not have. It is observed in (Bransford et al., 1999) that novices tend to solve problems concretely (plugging values into equations, e.g.), while experts tend to apply the (correct) abstract principles–it may be that novices cannot represent the abstract version of the task.

From the available data we can not determine which, if any, of these are true, but answering this is an interesting possible research topic.

Regarding the different amounts of doodling at the different institutions, there are a number of possible causes–local culture, the way programming is taught, the way the test was administered, chance, and so forth. It may simply be that the subject pools were fairly homogeneous within institutions.

## 8 Related work

Davies (1993) found that novices and experts have different ways of annotating programs, and also that experts spend more time annotating than novices. Both Davies' and our results indicate that annotations are a successful strategy for finding the correct answer.

Thomas et al. (2004) found that students who drew object diagrams performed better on tests involving object references, but their attempts to encourage greater use of diagrams were largely ineffective. Indeed, even after the benefits of using diagrams were demonstrated to students, they failed to use them when taking exams.

Hegarty (2004) looks at relations between externally produced diagrams and internal visualizations. One of the relations that she proposes is the use of external visualizations to augment internal visualizations: some of the internal processing and memory is "off-loaded" to an external representation. Experimental evidence in (Hegarty and Steinhoff, 1997) showed that "low-spatial" subjects who made annotations on an external diagram performed as well as "high-spatial" subjects when doing problems involving the inference of mechanical component motion, which suggests that the use of external annotation can substitute for keeping track of details internally.

Perkins et al. (1989) identify tracing (which they refer to as "close tracking") as a fundamental skill required by programmers: even for novices, it can be used to avoid, diagnose, and repair bugs in programs. They also observed that students fail to trace their code effectively, and identify a number of reasons for such failure:

1. students do not understand that tracing can be useful, or are not confident that they can trace effectivly,
2. students do not understand the programming language primitives,
3. students "project intentions" onto the code, and reason from these rather than from the code itself, and
4. students have differing cognitive styles.

We explicitly observed the third reason, students "recognizing" what a code fragment does and reasoning from that level, and suspect that the other three also occurred in this study.

More generally, many previous relevant studies, specifically regarding novice programmers, use of strategies, and distinguishing comprehension and generation, are cited in a survey by Robins et al. (2003).

## 9 Conclusions

This project used both fixed-code and skeleton-code questions to test concepts and identify misconceptions in introductory programming students. Performance in both improves when students annotate their tests, particularly by tracing on paper. There are differences, however: fixed-code questions require only understanding of how each expression works. Skeleton-code questions are closer to writing code, and can require more abstract reasoning. As they require code to meet a specification, they require students to reason about aggregate function, determine proper test cases, and so forth. Other strategies, such as process-of-elimination, may be more appropriate here, as the space of four or five different pieces of code and a number of possible test cases may make tracing too tedious.

In addition to the questions as to why students annotate harder questions less, this work suggests some areas for further study:

**Novice/expert annotation.** Do the annotation patterns used by individuals change over time? Some previous work (Davies, 1993) suggests both the number and the type of annotations differs between novices and experts. Can similar differences be seen between first and last year students and what would these differences mean?

**Institutional differences.** Are institutional differences as large as these data (Table 5) would indicate? Do these differences reflect differences in how programming is being taught, or culture, or simply differences in the way the data were collected?

**MCQs in practice.** The results in the paper shows a clear difference between fixed-code and skeleton-code questions which makes them appropriate for different stages in a first programming course. How might we best exploit these differences to improve the student learning experience?

## 10 Acknowledgments

Thanks to all of the working group participants (see Lister et al. (2004)), and the local arrangements people at Leeds who provided a great work space and plenty of tea and coffee. Thanks to the organizers, reviewers and participants of *Kolin Kolistelut*: the organizers who provided an intellectually stimulating environment, and the reviewers and participants who asked good questions and offered good ideas on how to improve this paper and build on these results. Finally, thanks to Sally Fincher, Marian Petre, and Josh Tenenberg, whose Bootstrapping and Scaffolding workshops (supported by NSF grants DUE-0122560 and DUE-0243242) had a large positive influence on this work.

## References

Boehm-Davis, D. A., 1988. Handbook of Human-Computer Interaction. Elsevier, Ch. 5 (Software Comprehension), pp. 107–121.

Bransford, J. D., Brown, A. L., Cocking, R. R. (Eds.), 1999. How people learn: Brain, mind, experience, and school. National Academy Press, Washington, D.C.

Davies, S., 1993. Externalising information during coding activities: effects of expertise, environment, and task. In: Empirical studies of programmers: 5th workshop. Ablex, Norwood, NJ, pp. 42–61.

Hegarty, M., 2004. Diagrams in the mind and in the world: relations between internal and external visualizations. In: Blackwell, A., Marriot, K., Shimojima, A. (Eds.), Diagrams 2004, LNAI 2980. Springer-Verlag, Berlin Heidelberg, pp. 1–13.

Hegarty, M., Steinhoff, K., 1997. Use of diagrams as external memory in a mechanical reasoning task. Learning and Individual Differences 9, 19–42.

Lister, R., Adams, E., Fitzgerald, S., Fone, W., Hamer, J., Lindholm, M., McCartney, R., Moström, J., Sanders, K., Seppälä, O., Simon, B., Thomas, L., 2004. A multi-national study of reading and tracing skills in novice programmers. SigCSE Bulletin (to appear).

Perkins, D. N., Hancock, C., Hobbs, R., Martin, F., Simmons, R., 1989. Conditions of learning in novice programmers. In: Soloway, E., Spohrer, J. C. (Eds.), Studying the Novice Programmer. Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp. 261–279.

Robins, A., Rountree, J., Rountree, N., 2003. Learning and teaching programming: a review and discussion. Computer Science Education 13 (2), 137–172.

Thomas, L., Ratcliffe, M., Thomasson, B., 2004. Scaffolding with object diagrams in first year programming classes: some unexpected results. In: Proceedings of the 35th SIGCSE technical symposium on computer science education. Norfolk, USA, pp. 250–254.