Do Computer Science Students Know What They Know?: A Calibration Study of Data Structure Knowledge

Laurie Murphy
Department of Computer Science
and Computer Engineering
Pacific Lutheran University
Tacoma WA 98447-0003 USA
lmurphy@plu.edu

Josh Tenenberg
Computing and Software Systems
Institute of Technology
University of Washington, Tacoma
Tacoma WA 98402-3100 USA
jtenenbg@u.washington.edu

ABSTRACT

This paper describes an empirical study that investigates the knowledge that Computer Science students have about the extent of their own previous learning. The study compares self-generated estimates of performance with actual performance on a data structures quiz taken by undergraduate students in courses requiring data structures as a pre-requisite. The study is contextualized and grounded within a research paradigm in Psychology called *calibration of knowledge* that suggests that self-knowledge across a range of disciplines is highly unreliable. Such self-knowledge is important because of its role in *meta-cognition*, particularly in cognitive self-regulation and monitoring. It is also important because of the credence that faculty give to student self-reports. Our results indicate that Computer Science student self-estimates correlate moderately with their performance on a quiz, more so for estimates provided after they have taken the quiz than before. The pedagogical implications are that students should be provided with regular opportunities for empirical validation of their knowledge as well as being taught the metacognitive skills of regular self-testing in order to overcome validation bias.

Categories and Subject Descriptors

K.3.2 [**Computers & Education**]: Computer & Information Science Education—*Computer Science Education*

General Terms

Human Factors

Keywords

Calibration of knowledge, data structures, metacognition, self-assessment

ITiCSE'05, June 27–29, 2005, Monte de Caparica, Portugal.

1. INTRODUCTION

Do computer science students know what they know? The answer to this question has several implications. Most importantly, students use self-knowledge for metacognitive control of their own learning, e.g. "overconfidence is a common phenomenon among young adult students that may result in inadequate learning due to premature termination of cognition" [13, p384]. Professors use student self-reports to make determinations about what to review at the start of a class, for example, when a CS professor asks herself, "Are my students weak in using linked lists?". Additionally, assessment of teaching effectiveness sometimes hinges on student self-reports as a measure of the effectiveness of a learning intervention, as in the claim "students reported that they improved as a result." But do these student self-assessments bear a close enough correspondence to more objective measures of performance to justify the confidence students and faculty sometimes place on them?

This paper presents an empirical study of student self-knowledge of data structures concepts. The study addressed the following specific research questions: Do students have systematic misconceptions or lack of retention concerning the data structures material in subsequent courses? Does student self-assessment correlate with performance? Is there a relationship between level of performance and level of self-knowledge? The discussion of research methods and results is preceded by a brief discussion of Psychology research in *calibration of knowledge* and followed by a discussion of pedagogical implications.

2. BACKGROUND

Self-assessment of knowledge is one form of *metacognition*, which Brown [2, p66] states "refers loosely to one's knowledge and control of [one's] own cognitive system". There have been a number of studies exploring the relationship of self-assessment of ability to performance across a number of domains, commonly referred to as studies of *knowledge calibration*. Performance is typically measured using a test of knowledge or ability [5, 12, 7], though course grades and evaluations by peers or supervisors are also used [14]. Self-assessments are commonly obtained by prompting subjects to estimate, for a given test, "how many test questions ... they thought they had answered correctly" [12, p1124].

The self-estimates of performance might be made before the exam, called *prediction*, after the exam, called *postdiction*, or both before and after. The main measure of sub-

ject calibration is the Pearson product-moment coefficient (r) between self-estimates and performance within a study population, though Goodman-Kruskal's gamma (G) (treating test scores as ordinal data) is sometimes used. If subjects within a population are well calibrated, we would expect correlations to be positive and close to 1, since this indicates that estimates and performance increase linearly. The research results do not present a clear and consistent picture, however. Correlations range from moderate negative correlation, $r = -.42, p < .001$ in [6], to non-significant correlations close to 0, $G = .02$ in [7] and $0.05 \leq r \leq 0.19$ in [12], to moderate positive correlation, r=0.46 in [5]. In a meta-analysis of 55 calibration studies with a combined population of 14,811 subjects across a wide variety of domains (e.g. clerical skills, managerial skills, college coursework, physical abilities, medical skills) Mabe and West [14] reported an overall correlation of $r = .29$.

Correlation alone, however, does not fully represent accuracy of self-estimates. As Kruger and Dunning pointed out [12, p190] "a high correlation between judgement and reality does not necessarily imply high accuracy." Another measure of calibration, though one rarely used in the calibration literature, examines the mean of the absolute magnitude of the difference between estimates and criterion scores, what we call *estimation error*. Fitzgerald et al carried out a calibration study with first-year medical students, where they simply added the one calibration question "Please estimate your percent correct on this exam (0% - 100%)" to each of the exams given in all first-year courses at the University of Michigan Medical School [6]. They concluded: "The high level of accuracy in these students' self-assessments (within 1% of their actual performance) is striking, and suggests well-developed self-assessment skills." Kruger and Dunning reported mean estimation errors of 3.48 and 1.84 in the bottom and top performance quartiles, respectively in a study in which subjects graded five 20-item exams of other students (study 3, phase 2, in [12]). This error represents the difference between the number of problems graders scored as correct and the actual number of problems correct. Lin and Zabrucky reported in [13] on a study by Glover [8], in which student subjects had mean estimation errors (predicted to actual) of 1.21 and 7.43 in the bottom and top performance quartiles, respectively, on the Nelson-Denny Reading Test.

Researchers have attempted to determine if different subject popluations have different calibration ability. Rammstedt and Rammsayer investigated the effect of gender on calibration and found [16, p869] that "there was some direct evidence for the assumption that estimates of intelligence are susceptible to gender stereotypes." Ackerman et al found that students with college majors in the Social Sciences or Humanities were accurately calibrated across a variety of knowledge domains, whereas Business majors consistently over-estimated performance across all domains [1].

Several researchers have attempted to determine if subject domain expertise has a bearing on calibration accuracy. Lin and Zabrucky [13], as well as Fitzgerald *et al.* [6] cited several studies that provide evidence that those with high domain expertise often have the "illusion of knowing": more knowledge sometimes brings along with it a sense of overconfidence. This is closely related to the phenomenon of *validation bias* summarized in [4, p6]: "People do not dispassionately count up their success and failures to form a self-impression as much as they actively interpret them to fit chronic views, usually positive ones, of the self ... Positive feedback is more likely to be accepted unquestioningly; negative feedback is placed under close scrutiny with an eye toward discounting it." In contrast to over-confident estimates being associated with high expertise, Kruger and Dunning suggested that over-confidence is associated with low expertise [12, p1121]: "when people are incompetent in the strategies they adopt to achieve success and satisfaction, they suffer a dual burden: Not only do they reach erroneous conclusions and make unfortunate choices, but their incompetence robs them of the ability to realize it. ... In essence ... the skills that engender competence in a particular domain are often the very same skills necessary to evaluate competence in that domain." In studying the calibration of college students on tests in humor, logic, and grammar, they found considerable over-confidence among bottom quartile performers and a small amount of under-confidence among top quartile performers. Krueger and Mueller [11, p184] argued that this relationship between estimates and performance stems not from any metacognitive differences between expertise-based groups, but from a general tendency to overestimate performance combined with the statistical artifact of *regression toward the mean*. "With repeated testing, high and low test scores regress toward the group average, and the magnitude of these regression effects is proportional to the size of the error variance and the extremity of the initial score."

To summarize, a number of studies indicate moderate calibration ability by students. Subject population characteristics, such as gender and college major, appear to influence calibration ability, while the evidence is conflicting on whether domain expertise is related to calibration ability.

3. STUDY METHODOLOGY

The study examined upper-level computer science students' ability to self-assess their prerequisite data structures knowledge. As described below, students from two universities took a quiz to measure their data structures knowledge and completed both pre- and post-quiz self-assessment questionnaires to determine their calibration ability. The research protocol, quiz and self-assessment questionnaires were approved by the Institutional Review Boards at both the University of Washington, Tacoma and Pacific Lutheran University.

3.1 Subjects

Participants were 61 undergraduate students enrolled in upper-level computer science classes at two universities in the Pacific Northwest of the USA. Twenty-eight subjects were from Pacific Lutheran University (PLU), a private, suburban, liberal arts university. Thirty-three were from the University of Washington, Tacoma (UWT), a public, urban university serving junior and senior level students, the majority of whom are community college transfers. Seventy-eight students enrolled in four targeted classes were given full credit for completing the data structures quiz, regardless of their score. Only quiz results from 61 students giving their consent are included in this study. Data on gender was not collected due to low enrollment of female students.

3.2 Targeted Classes

The study was conducted in four upper-level undergraduate classes, two at each institution. These included *Programming Languages* and *Algorithms* at PLU, and *Algorithms* and *Software Engineering* at UWT. These classes were selected because they are required courses for all computer science majors at their respective institutions, and they require data structures as a prerequisite. Three of the courses have additional prerequisites: both algorithms courses also require *Discrete Math*, and the software engineering course requires both *Technical Team Management* and *Algorithms*.

3.3 Data Structures Prerequisite

The data structures courses at PLU and UWT differ in terms of whether a closed lab is associated, whether the course is quarter or semester, and the amount of time devoted to specific topics such as graphs and object orientation. There is, however, significant overlap. Both institutions include the study of fundamental data structure abstractions and implementations including lists, stacks, queues, trees and hash tables. Both also cover recursion and algorithm analysis, particularly within the context of sorting and searching. Additionally, both courses serve as the typical "gateway" prerequisite to most upper-level computer science classes. The similarities of context enabled the same quiz to legitimately be given at both institutions; the differences of context increase confidence in the generalizability of the results beyond the students at either institution.

3.4 Quiz Construction

To assess students' data structures knowledge, we constructed a quiz[1] using multiple-choice questions from Advanced Placement (AP) and Graduate Record Examination (GRE) computer science practice tests [3, 10, 9, 17]. The AP and GRE questions increased external validity and reduced bias in favor of students at either university. The multiple-choice format also provided unambiguous correct answers and allowed us to accurately gauge the number of questions we could reasonably expect students to answer in 30 minutes. Questions were selected to closely reflect the topics covered in a typical data structures course. They were also reviewed by the primary data structures instructor at each institution for consistency with their course syllabi, especially the proportion of questions on the different topics (see Table 1 for question topics and bibliographic source for each question). Furthermore, to confirm the questions and time constraints were fair and reasonable for our subjects, we also piloted the quiz with four upper-level computer science majors, two from each institution.

3.5 Procedure

The quiz was administered in class during the first week of the fall 2003 term. Students were informed both verbally and in writing that they were required to take the unannounced quiz, but that they would be given full credit for taking it, regardless of their scores.

In addition to the quiz, students completed both pre- and post-quiz questionnaires to assess their calibration ability. To enable students to make an accurate prediction of their performance on the quiz, we provided them with the following detailed description on the pre-quiz questionnaire:

> You will be completing a 14-question multiple-choice quiz covering material that was presented in your **Data Structures** course. The questions are primarily taken from College Board Advanced Placement (AP) practice books. In particular, this quiz will test your knowledge of *trees*, *linked lists*, *stacks*, *queues*, and *hash tables*. For each of these topics, there may be questions concerning data structure definitions, operations, implementations, worst-case time analysis, and trade-offs between different data structure choices. In addition, there will be questions about different *sorting* and *searching* algorithms.

Students predicted the absolute number of questions they would answer correctly by responding to the following:

> Based on your assessment of your knowledge of the data structures material, how many questions in the 14-question quiz do you predict you will get correct?

We also asked students to rate their score on a percentile basis relative to other students taking the quiz, and their level of difficulty with and interest in the data structures material. The analysis discussed in this paper does not include results from relative estimations, interest, or difficulty. Full details can be found in [15].

4. RESULTS AND DISCUSSION

4.1 Data structure knowledge

Performance by students from PLU (N = 28, M = 8.36, SD = 2.48) was virtually identical to that of students from UWT (N = 33, M = 8.42, SD = 2.59), and an independent groups t test indicated no significant difference ($t(59) = -.103$, $p = 0.92$) between these groups. For the balance of this paper, all students will be treated as belonging to the same population.

The mean score for the population of students was just above one-half of the questions (N = 61, M = 8.39, SD = 2.52), with one student scoring the maximum possible score of 14 and five students with the lowest score of 4. There is probably some upward bias in these results, since students who did not give consent to use their quiz results were primarily those who had dropped or were doing poorly.

Table 1 shows both the number and the percentage of subjects answering each question correctly. Students performed best on questions testing knowledge of the stack, queue, and tree interface, and performed worst on questions testing knowledge of comparing runtime effeciency of binary and sequential search, as well as in identifying whether a piece of code is an example of selection, insertion, mergesort, or quicksort. Questions involving code tracing or implementation knowledge of lists, trees, and recursion were answered correctly by one-half to two-thirds of the students.

4.2 Knowledge calibration

Descriptive statistics are provided for the full sample of 61 students in Table 2 for actual score, prediction, postdiction, prediction error, and postdiction error. We defined prediction error as the absolute value of the difference between a subject's predicted and actual score, similarly for postdiction error.

[1]This use of a composite exam is consistent with the Fair Use Statute of Section 107 of the US Copyright Act of 1976.

Table 1: Subjects answering correctly (N = 61)

Topic [Source]	Subjects answering correctly	
singly linked list properties/analysis [9, p16]	35	57%
stack vs queue: choosing right DS [10, p248]	52	85%
binary vs sequential search [10, p161]	19	31%
binary search tree properties [10, p255]	54	89%
binary tree traversals [3]	50	82%
hash table properties/definitions [10, p305]	27	44%
sorting: merge vs insertion sort [17, p310]	41	67%
bst insertion/traversal + analysis [10, p263]	37	61%
tracing recursive binary tree methods [3]	28	46%
recursive binary tree method analysis [3]	29	48%
tracing stack operations [3]	57	93%
sorting algorithm identification [17, p310]	13	21%
singly linked list traversal [3]	39	64%
singly linked list traversal analysis [3]	31	51%

Table 2: Descriptive statistics

	Min	Max	M	SD
Actual score	4	14	8.39	2.52
Prediction	5	14	9.69	2.08
Postdiction	4	14	8.99	2.41
Prediction error (|Actual score - Prediction|)	0	7	2.30	1.62
Postdiction error (|Actual score - Postdiction|)	0	7	1.65	1.40

Paired samples t tests indicated that the difference between mean actual scores and predictions is significant ($t(60) = -4.03$, $p < .001$), as is the difference between mean actual scores and postdictions ($t(60) = -2.24$, $p < .05$).

Table 3 shows the correlations (Pearson's product-moment coefficient, r) between estimations, actual score, and estimation error. Overall, both predictions and postdictions are positively and significantly correlated with actual scores. Since the predictions were made prior to viewing the exam questions, they are based on more generalized student beliefs about their data structures knowledge, cued by the topic areas specified in the directions (e.g. linked lists, trees). It is not surprising then, that both correlation increases and estimation error descreases after subjects view the exam itself.

Overall, prediction calibration is moderate, with postdiction calibration being relatively high, especially in comparison to studies cited above. We believe this high postdiction calibration is a result of several factors. One is that much of computer science in general, and data structures in particular, lends itself to high calibration given its objective nature.

Table 3: Pearson's Product-Moment Coefficient

	Actual score	Pred.	Postd.	Pred. error
Pred.	.418**			
Postd.	.643**	.581**		
Pred. error	-.464**	.158	-.145	
Postd. error	-.249	-.011	.055	.378**

** Correlation is significant at the 0.01 level (2-tailed)

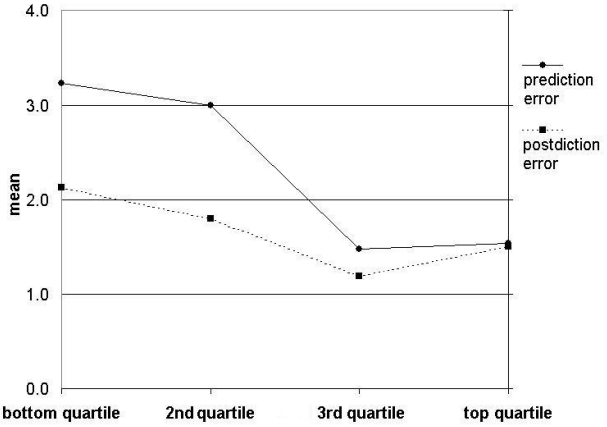

Figure 1: Raw Score Error by Quartile

Second, we took care to use clearly stated questions having definitive answers, since, as the literature indicates, low estimation accuracy might simply reflect ambiguity in exam questions or instructions to the subjects. And third, consistent with Mabe and West's [14] findings that calibration improves when subjects expect self-estimates to be validated, the setting of the exam made clear that self-estimates would be compared to actual performance, thus reducing some of the incentives to inflate estimates.

Did those performing the worst provide the least accurate predictions? Table 3 shows that there is a moderate, inverse correlation between prediction and calibration error, i.e. error decreases as scores increase. But the negative correlation between error and performance is weak and non-significant for raw score postdiction.

Figure 1 provides a more detailed view of error by quartile. The shape of these error curves provides some evidence for regression toward the mean, though the poor calibration of the second quartile and the asymmetry in error magnitude between the bottom and top quartiles indicates that this alone does not account for error.

What the correlation and error statistics do indicate is that there is not in general a double burden for the lowest performers. Though their calibration accuracy was less than that of the highest performers, the bottom quartiles also improved the most in going from prediction to postdiction, hence displaying the sort of metacognitive estimate of performance that they could use to regulate their study. If there are lessons here concerning metacognition, it might be that lower performers overestimate their general abilities (what prediction estimates are presumably based on), but more accurately calibrate following direct experience. Interestingly, the estimates of students performing in the top quartile remained virtually unchanged in going from prediction to postdiction. In neither test did top quartile students on average overestimate their scores and display the "illusion of knowing" that is often associated with performances that subjects find relatively easy.

5. CONCLUSION

We set out to answer a number of specific research questions about student knowledge and metaknowledge in computer science: *Do students have systematic misconceptions*

or lack of retention concerning the data structures material in subsequent courses? Quiz results revealed particular areas of weakness, including sorting and searching algorithms, hash tables, and analyzing recursive binary tree methods. However, it is unclear whether the weaknesses were due to insufficient learning or lack of retention. *Does student self-assessment correlate with performance?* Computer science students' calibration accuracy was generally higher than that of students in other fields, possibly because of their familiarity with the data structures material or the objective nature of the domain. Results also revealed a moderate correlation between student estimates and their quiz performances, more so for estimates provided after they had taken the quiz than before. *Is there a relationship between level of performance and level of self-knowledge?* Students with higher scores had better calibration ability, and did not display the *illusion of knowing* by making overconfident estimates of performance. Students with lower scores did not appear to suffer the *dual burden of incompetence*, showing the most improvement from prediction to postdiction of raw scores.

Our results lead to several pedagogical implications. Since student estimates of performance prior to or in place of taking an exam offer somewhat inaccurate assessments of actual knowledge, such predictions should be used cautiously as gauges for faculty action (such as topics to review). Although the top performing students were well calibrated both before and after an exam, the weakest students–those requiring the most remediation–are the ones most likely to make inaccurate predictions. Even if reviews are targeted appropriately, students might still rely on domain familiarity or estimates of general knowledge to systematically ignore topics of weakness. Therefore, it is essential to provide students with explicit opportunities to assess prerequisite knowledge. Objective assessments of ability may be more useful than subjective measures because they make negative feedback more evident and difficult to discount. As research on calibration of comprehension has suggested, "self-generated feedback has a more positive impact on calibration than does other-provided feedback" [13, p384]. Instructors who encourage students to use regular self-assessments of performance, such as practice exams, thus provide students with learning practices that can help them to overcome their own validation biases.

6. ACKNOWLEDGMENTS

We are grateful to David Wolff and Donald Chinn for providing feedback on the quizzes and insight into the data structures courses that they teach at PLU and UWT; to Karen Furuya, Dan Bahrt, Kenneth Keeler, and Darrel Rohar for piloting versions of the quiz; to Bronwyn Pughe for editorial assistance, and to Sally Fincher, Marian Petre, and the participants of the *Bootstrapping Research in Computer Science Education* project for feedback and encouragement to carry out this study. This material is based upon work supported by the National Science Foundation under Grant No. DUE-0122560. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7. REFERENCES

[1] P. L. Ackerman, M. E. Beier, and K. R. Bowen. What we really know about our abilities and our knowledge. *Personality and Individual Differences*, 33:587–605, 2002.

[2] A. Brown. Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. Weinert and R. Kluwe, editors, *Metacognition, Motivation, and Understanding*, pages 65–116. Lawrence Erlbaum Associates, Inc., 1987.

[3] *The College Board AP Advanced Placement Program Computer Science Course Description*. Retrieved 20 July 2003, from `www.collegeboard.com`, 2003.

[4] J. Ehrlinger and D. Dunning. How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology*, 84(1):5–17, 2003.

[5] H. T. Everson and S. Tobias. The ability to estimate knowledge and performance in college: A metacognitive analysis. *Instructional Science*, 26:65–79, 1998.

[6] T. Fitzgerald, L. Gruppen, C. White, and W. Davis. Medical student self-assessment abilities: Accuracy and calibration. In *Annual Meeting of the American Educational Research Association*, 1997.

[7] A. M. Glenberg and W. Epstein. Inexpert calibration of comprehension. *Memory and Cognition*, 15(1):84–93, 1987.

[8] J. A. Glover. Improving readers' estimates of learning from text: The role of inserted questions. *Reading Research Quarterly*, 28:68–75, 1989.

[9] *Graduate Record Examinations Computer Science Test Practice Book*. Retrieved 20 July 2003, from `www.gre.org`, 2001.

[10] S. Horowitz. *Review for the Computer Science AP Exam in C++, 2e*. Addison-Wesley Longman, Inc., 2000.

[11] J. Krueger and R. A. Mueller. Unskilled, unaware, or both: The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, 82(2):180–188, 2002.

[12] J. Kruger and D. Dunning. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6):1121–1134, 1999.

[13] L. M. Lin and K. Zabrucky. Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology*, 23:345–391, 1998.

[14] P. A. Mabe III and S. G. West. Validity of self-evaluation of ability: A review and meta-analysis. *Applied Psychology*, 67(3):280–296, 1982.

[15] L. Murphy and J. Tenenberg. Knowing what I know: an investigation of undergraduate knowledge and self-knowledge of Data Structures. Forthcoming in *Computer Science Education*, 15(4), 2005.

[16] B. Rammstedt and T. Rammsayer. Sex differences in self-estimates of different aspects of intelligence. *Personality and Individual Differences*, 29:869–880, 2000.

[17] R. Teukolsky. *How to Prepare for the AP Computer Science Advanced Placement Examination*. Barron's Educational Series, 2001.