Computer Science Education
Vol. 15, No. 4, December 2005, pp. 297–315

Routledge
Taylor & Francis Group

Knowing What I Know: An investigation of undergraduate knowledge and self-knowledge of data structures

Josh Tenenberg*[a] and Laurie Murphy[b]
[a]*Computing and Software Systems, Institute of Technology, University of Washington, Tacoma, Tacoma WA 98402-3100, and* [b]*Department of Computer Science and Computer Engineering, Pacific Lutheran University, Tacoma WA 98447-0003*

This paper describes an empirical study that investigated the knowledge that Computer Science students have about the extent of their own previous learning. The study compared self-generated estimates of performance with actual performance on a data structures quiz taken by undergraduate students in courses requiring data structures as a prerequisite. The study was 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, as well as in the credence that instructors give to student self-reports. Our results indicated that Computer Science student self-estimates are highly correlated with performance, more so for estimates provided after the performance than before. This high level of calibration, however, was likely the result of a number of conditions that do not always hold: that the students already had domain expertise, that the quiz had unambiguous and verifiable answers, and that students expected their estimates to be validated. When these conditions are not met, it becomes more important for students to have direct feedback about their performance so as to uncover those areas where their intuitions might mislead them. Students also had weak knowledge about their standing relative to their peers, particularly those in the lower performance quartiles, exhibiting the well known *better-than-average* heuristic. There was, additionally, no correlation between calibration ability and degree of liking or difficulty with the data structures material, suggesting that instructors and researchers should not treat liking or difficulty as reliable indicators of the learning that has occurred.

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

*Corresponding author. Computing and Software Systems, Institute of Technology, University of Washington, Tacoma, Tacoma WA 98402-3100. E-mail: jtenenbg@u.washington.edu

ISSN 0899-3408 (print)/ISSN 1744-5175 (online)/05/040297-19 © 2005 Taylor & Francis
DOI: 10.1080/08993400500307677

metacognitive control of their own learning. For example, Lin and Zabrucky (1998, p. 384) report that young adults commonly terminate cognition prematurely due to overconfidence. Instructors use student self-reports in choosing what to review at the start of a class, for example when asking, ''Are students weak in using linked lists?''. Similarly, assessments of teaching effectiveness sometimes hinge on student self-reports, 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 instructors sometimes place on them? This paper provides a brief summary of research in *calibration of knowledge*, where self-assessment of ability is compared to measures of performance. Following this, we ask a number of specific research questions related to teaching Computer Science which can be informed by empirical evidence. We then discuss an empirical study that we undertook in which we measured student self-assessment of data structure knowledge against their performance on a data structures exam. We report on both student test performance, as well as on the relationship between their self-assessments and performance. Finally, we discuss the implications of research in knowledge calibration and our results in particular, for teaching computer science.

2. BACKGROUND

Self-assessment of knowledge is one form of *metacognition*, which Brown (1987) states ''refers loosely to one's knowledge and control of [one's] own cognitive system'' (p. 66). 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 (Everson & Tobias, 1998; Kruger & Dunning, 1999; Glenberg & Epstein, 1987), though course grades and evaluations by peers or supervisors are also used (Mabe & West, 1982). Self-assessments are commonly obtained by prompting subjects to estimate how well they will perform on a given test of ability. Self-estimates that are made after a performance are called *postdictions*, and reflect both knowledge about performance and about the task itself. Self-estimates that are made before a performance are called *predictions*. Predictions often require estimates not only of performance but of the specifics of the performance task that might be unknown until after the task has been performed. This would be the case for students who are asked to predict the number of questions that they expect to answer correctly on an exam that they have not yet seen. It is therefore unsurprising that several studies demonstrate that postdiction is generally more accurate than prediction (Mabe & West, 1982; Lin & Zabrucky, 1998).

The main measure of subject 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 Fitzgerald et al. (1997), to non-significant correlations close to 0, $G = .02$ in Glenberg and Epstein (1987) and $0.05 \leq r \leq 0.19$ in Kruger and Dunning (1999), to moderate positive correlation, $r = 0.46$ in Everson and Tobias (1998). 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 reported an overall correlation of $r = .29$.

Correlation alone, however, does not fully represent accuracy of self-estimates, as Kruger and Dunning (1999, p. 190) point out. Another measure of calibration that is sometimes used examines the mean of the absolute magnitude of the difference between estimates and criterion scores, what we call *estimation error*. To see how measures of estimation error can provide a more nuanced view of study data than correlation alone, consider Table 1, which shows hypothetical calibration data for two different subject samples. The actual scores are identical for both samples, and the Pearson correlation coefficient for both is significant and close to 1. But the samples differ dramatically in terms of estimation error, with the first sample having an average error of 24 and the second sample having an average error of 1.

Fitzgerald et al. (1997) 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. "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 (1999, study 3, phase 2) 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

Table 1. Correlation and mean estimation error

	Sample 1		Sample 2	
	Actual	Estimate	Actual	Estimate
	30	0	30	29
	40	10	40	39
	50	50	50	51
	60	90	60	61
	70	100	70	71
mean	50	50	50	50.2
r	.978**		.999**	
mean | estimation – actual |	24		1	

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

20-item exams of other students. This error represents the difference between the number of problems graders scored as correct and the actual number of problems correct. Lin and Zabrucky (1998) reported on a study by Glover (1989), 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.

Mabe and West (1982) suggested that people are in general better at self-assessing their abilities *relative* to others than on an absolute scale. ''The issue generally is not one of how much ability a person has in the absolute sense but rather of how much ability he or she has in comparison with other people.'' (p. 290) They found that the use of social comparison measures (e.g. ''as compared to other students in your class'') was one of the main situational factors in the 55 calibration studies that they examined significantly correlated with high calibration ability.

Kruger (1999), however, cited evidence from a number of empirical studies indicating that people systematically over-estimate their abilities relative to others: ''As the author of one of the best-selling introductory psychology textbooks in the United States put it, 'for nearly any subjective and socially desirable dimension . . . most people see themselves as better than average' (Myers, 1998, p. 440).'' (p. 221) This is often referred to as the *better-than-average* or *Lake Wobegone effect*.

Kruger and Dunning (1999) suggested that systematic over-estimation of relative performance is associated with low expertise: ''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.'' (p. 1121) In studying the calibration of college students on tests in humor, logic, and grammar, they found considerable over-estimation of performance relative to their university classmates among bottom quartile performers and a small amount of under-estimation among top quartile performers. Krueger and Mueller (2002) argued that this relationship between estimates of relative ranking and actual performance does not stem from any metacognitive differences between expertise-based groups. Rather, it results from subjects generally employing a better-than-average heuristic along 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.'' (p. 184)

In an attempt to determine the factors that influence calibration differences, researchers have both experimentally manipulated measurement conditions as well as asked subjects to self-rate along a number of dimensions other than direct ratings of exam performance. These dimensions included asking subjects to self-rate their domain knowledge (e.g. Chemistry or Electronics, Ackerman et al., 2002), general ability (e.g. ''general logical reasoning ability'', Kruger and Dunning (1999), or ''verbal ability'', Ackerman et al. (2002)), affective assessment of the domain

(e.g. desirability or interest of domain investigated, Lin and Zabrucky, 1998), or subjective assessment of subject difficulty (e.g., Kruger, 1999).

Kruger (1999) reported systematic under-estimation on social comparison measures in domains that subjects rate as difficult and over-estimation in domains that subjects rate as easy. In partial disagreement with this result, Lin and Zabrucky (1998) concluded "according to available data from the postdiction paradigm, students' competence in past performance assessments varies as a function of test item difficulty. Students tend to exhibit less illusion of knowing on easy test items and are able to calibrate performance on those items with greater accuracy than on difficult items" (p. 370). Mabe and West (1982, p. 293) reported from their meta-analysis of 55 calibration studies that "expectation of self-evaluation validation" by subjects is one of the key factors in prediction accuracy. And Ackerman et al. (2002) found that calibration accuracy with the same population of college graduates from a variety of majors differs dramatically depending on subject domain. "The largest correlations [between self-assessments and performance] were found for the Science domains (mean $r = 0.52$), followed by Civics (mean $r = 0.45$), Humanities (mean $r = 0.45$) and Business (mean $r = 0.16$)" (p. 599).

Lin and Zabrucky (1998) conjectured that "individual factors, such as interest and motivation, may mediate the degree to which readers can precisely judge whether a text is fully comprehended." (p. 363) However, they reported that there is limited evidence for this, with one study showing a small correlation between interest and calibration ($r = .15$)

Researchers have also attempted to determine if different subject populations have different calibration ability. Rammstedt and Rammsayer (2000) investigated the effect of gender on calibration and found that "there was some direct evidence for the assumption that estimates of intelligence are susceptible to gender stereotypes." (p. 869) Ackerman et al. (2002) 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.

Several other studies have also attempted to determine if subject domain expertise has a bearing on calibration accuracy. Lin and Zabrucky (1998), as well as Fitzgerald et al. (1997), 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. Agnew et al. (1994) cited a number of studies demonstrating validation bias associated with expert judgments. As Ehrlinger and Dunning (2003) pointed out, "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." (p. 6)

To summarize, a number of studies indicate moderate calibration ability by students, although several factors mediate this. These factors likely include characteristics of the subject domain, test item difficulty, student affect toward the

subject area, and conditions of the research study itself, such as whether subjects believed that estimates would be validated through performance tests. Subject population characteristics, such as gender and college major, also appear to influence calibration ability, while the evidence is conflicting on whether domain expertise is related to calibration ability.

3. RESEARCH QUESTIONS

We can now ask a number of specific questions regarding student knowledge and metaknowledge in computer science: Do students have systematic misconceptions or lack of retention concerning the data structures material in subsequent courses? Do student self-estimates correlate with performance, or do students systematically under- or over-estimate performance? Do most students rate themselves above average relative to their peers? Does self-assessment of subject difficulty or interest correlate with either performance or accuracy? Is calibration accuracy related to domain expertise? Do students with higher scores have the illusion of knowing? Do those with lower scores suffer the dual burden of incompetence?

We undertook the following empirical study to obtain evidence to inform these questions. As far as we are aware, these are the first studies to date reported in the scientific literature on calibration of knowledge within Computer Science. Further, there currently exists no data on student knowledge retention of data structure knowledge downstream from the introductory data structures course. Although this latter question is not our primary concern, it serves as an excellent test-bed for calibration research, especially given the relative standardization of the material in the data structures course.

4. STUDY METHODOLOGY

The study examined upper-level computer science students' ability to self-assess their prerequisite data structures knowledge. As described in more detail 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 (IRB's) at both the University of Washington, Tacoma and Pacific Lutheran University.

4.1. Subjects

Participants were drawn from 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. Approximately 70% of students enter PLU directly from high school and entering freshmen typically have a 3.0 or better high school grade point average on a scale from 0 to 4 and a 1000 out of 1600 on the College

Board SAT exam. Students transferring from other colleges or universities must usually have a grade point average of 2.5 or better. Thirty-three subjects were from the University of Washington, Tacoma (UWT), a public, urban university serving junior and senior level students. Greater than 90% of the students entering the computer science major transfer from community colleges, all entering students must have a cumulate grade point average from all previous courses of 2.0 on a scale from 0 to 4, and approximately two-thirds of the students entering the computer science major complete their degree. Most students were taught Java as their introductory language, and Java is used as the predominant language throughout the PLU and UWT computer science curricula.

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 enrolment of female students (as few as two in some targeted classes).

4.2. 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*.

4.3. Data Structures Prerequisite

There are distinct differences in the two prerequisite data structures courses. At PLU, the class is taught on a semester schedule, requires one semester of Java programming as a prerequisite, and has an accompanying closed lab. The UWT class is taught on quarters, requires two quarters of Java programming as prerequisites and does not have a closed laboratory component. The courses serve different types of students at slightly different points in their respective programs, and there is some dissimilarity in course content. For example, PLU focuses more on object oriented programming techniques and UWT covers a few more advanced topics, such as balanced trees and graphs. Despite their differences, the classes are largely similar in their traditional data structures content and in the primary role they play in their respective curricula. Classes at both institutions include the study of fundamental data structure abstractions and implementations including lists, stacks, queues, trees, and hash tables. They 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. These considerable similarities enabled the same data structures quiz to legitimately be administered to students at both

institutions; the differences in context increase confidence in the generalizability of the results beyond the students at either institution.

4.4. 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 (College Entrance Examination Board, 2003; Horowitz, 2000; Graduate Record Examination, 2001; Teukolsky, 2001). AP and GRE questions ensured external validity and prevented 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 2 for question topics and a 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.

4.5. 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

Table 2. Quiz Question Topics

Question	Topic & Source
1	singly linked list properties and analysis of ops[a(p:60)]
2	stacks vs. queues – choosing right data structure[b(p:248)]
3	binary search vs. sequential search[b(p:161)]
4	binary search tree general properties[b(p:222)]
5	binary tree traversals[c]
6	hash table properties/definitions[b(p:305)]
7	sorting – merge sort vs. insertion sort[d(p:310)]
8	BST insertion and traversal w/ analysis[b(p:263)]
9	tracing recursive binary tree methods[c]
10	analysis of recursive binary tree methods[c]
11	tracing stack operations[c]
12	sorting algorithm identification[d(p:310)]
13	singly linked list traversal[c]
14	singly linked list traversal analysis[c]

[a]Graduate Record Examination (2001), [b]Horowitz (2000), [c]College Entrance Examination Board (2003), [d]Teukolsky (2001)

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 tradeoffs between different data structure choices. In addition, there will be questions about different *sorting* and *searching* algorithms.

Students predicted both the absolute number of questions they would answer correctly and their relative percentile ranking compared to other students taking the quiz by responding to the following questions:

1. 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?
2. Compared to the other students taking this quiz, how do you think that you will place? Provide a number between 0 and 100 that indicates the percentage of students that you will perform better than. For example, ''10'' means that you will perform better than 10 percent of the students, and ''90'' means that you will perform better than 90 percent of the students.

To assess the influence individual factors, such as perceived relative domain difficulty and interest (Lin and Zabrucky, 1998, p. 363), have on performance or calibration, students were asked to give Likert-type responses to the following questions:

3. Rank the level of interest you have in the data structures material. (very uninteresting, somewhat uninteresting, neutral, somewhat interesting, very interesting)
4. Compared to the other courses that you have taken in college, rank the level of difficulty that you had in learning the data structures material. (very difficult, somewhat difficult, neutral, somewhat easy, very easy)

After completing the quiz in the allotted 30 minutes, students completed a second questionnaire on which they postdicted their absolute and relative quiz performance.

5. RESULTS AND DISCUSSION

5.1. 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 ($t(59) = 0{:}103$, $p = 0{:}92$) between these groups. Students in the more advanced course for which both Algorithms and Data Structures are prerequisites performed similarly to the population as a whole ($N = 17$, $M = 8.6$, $SD = 2.7$). 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 largely those who had dropped or were doing poorly in their classes.

Figure 1 shows, for each question, both the number and the percentage of subjects answering that 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 efficiency of binary and sequential searches, as well as in identifying whether a piece of code is an example of selection sort, insertion sort, mergesort, or quicksort. Questions involving code tracing or implementation knowledge of linked lists, trees, and recursion were answered correctly by one-half to two-thirds of the students.

5.2. Knowledge Calibration

5.2.1. Raw score error and correlation. Descriptive statistics are provided for the full sample of 61 students in Table 3 for actual score, prediction, postdiction, prediction

Question Number	Subjects Answering Correctly		Topic
1	35	57%	singly linked list properties and analysis of ops
2	52	85%	stacks vs. queues - choosing right data structure
3	19	31%	binary search vs. sequential search
4	54	89%	binary search tree general properties
5	50	82%	binary tree traversals
6	27	44%	hash table properties/definitions
7	41	67%	sorting - merge sort vs. insertion sort
8	37	61%	BST insertion & traversal w/ analysis
9	28	46%	tracing recursive binary tree methods
10	29	48%	analysis of recursive binary tree methods
11	57	93%	tracing stack operations
12	13	21%	sorting algorithm identification
13	39	64%	singly linked list traversal
14	31	51%	singly linked list traversal analysis

Figure 1. Subjects answering correctly ($N = 61$)

error, and postdiction error. We defined the prediction error of a subject as the absolute value of the difference between the subject's predicted and actual scores, similarly for postdiction error. Although this provides some measure of accuracy, it can underestimate metacognitive error. This would occur when a student believes that one question is answered correctly and another is answered incorrectly, both of which are false beliefs.

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 4 shows the correlations (Pearson's product-moment coefficient, r) between estimations, actual scores, and estimation error. Overall, both predictions and postdictions were 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 decreases after subjects view the exam itself.

Overall, prediction calibration is moderate, with postdiction calibration being relatively high, especially in comparison to research studies cited above. We believe that 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. Second, we took care in our study design to

Table 3. Descriptive statistics: Raw scores

	Min	Max	M	SD
Raw score actual	4	14	8.39	2.52
Raw score prediction	5	14	9.69	2.08
Raw score postdiction	4	14	8.99	2.41
Score prediction error (|Raw score − Prediction|)	0	7	2.30	1.62
Score postdiction error (|Raw score − Postdiction|)	0	7	1.65	1.40

Table 4. Pearson's Product-Moment Coecient: Raw scores

	Actual Score	Predicted score	Postdicted score	Prediction error
Predicted score	.418**			
Postdicted score	.643**	.581**		
Prediction error	−.464**	.158	−.145	
Postdiction error	−.249	−.011	.055	.378**

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

use clearly stated questions having definitive answers, since, as the literature indicates, low estimation accuracy might simply reflect ambiguity in exam questions or in the instructions to the subjects. And third, consistent with the findings of Mabe and West (1982) that calibration improves when subjects expect self-estimates to be validated, the setting of the exam made clear that subject self-estimates would be compared to actual performance, thus reducing some of the incentive to inflate estimates.

5.2.2. Calibration and expertise. Did those performing the worst provide the least accurate predictions? Table 4 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 2 provides a more detailed view of error by quartile. 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 over-estimate 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.

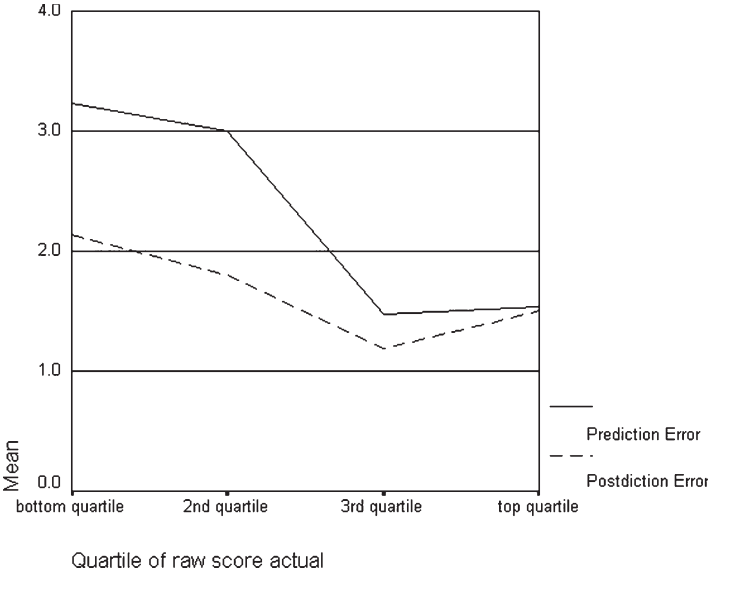

Figure 2. Raw score error by quartile

5.2.3. Interest and difficulty. As for affective and subjective factors, neither student interest in the data structures material nor ratings of subject difficulty showed any effect on knowledge calibration. Pairwise correlations using Kendall's τ (due to the ordinal nature of the data) on interest, difficulty, raw score prediction error, and raw score postdiction error did not differ significantly from 0. Furthermore, as shown in Figure 3, there was little difference in these ratings across performance groups.

5.2.4. Calibration on relative rankings. Over the entire population, subjects were moderately calibrated in their estimates of percentile ranks relative to their classmates as compared to their actual rankings based on quiz score. Pearson r coefficients of prediction to actual rank and postdiction to actual rank were 0.38 and 0.44, respectively, both significant at the 0.01 level (2-tailed). A closer look at the data, however, revealed that the estimation errors in ranking were not uniformly distributed across the population, with those in the lower two performance quartiles having considerably higher estimation errors than those in the top half, as illustrated in Figure 4. Pearson r coefficients of prediction error to actual rank and postdiction error to actual rank were -0.65 and -0.43, respectively, both significant at the 0.01 level (2-tailed). That is, as actual scores decrease, estimation errors increase. The raw data itself revealed that only 7 of 61 subjects predictively and 11 of 61 subjects postdictively ranked themselves in the bottom two quartiles. Taken in total, these results were consistent with the hypothesis that subjects will, in general, rank themselves as better than average in social comparisons.

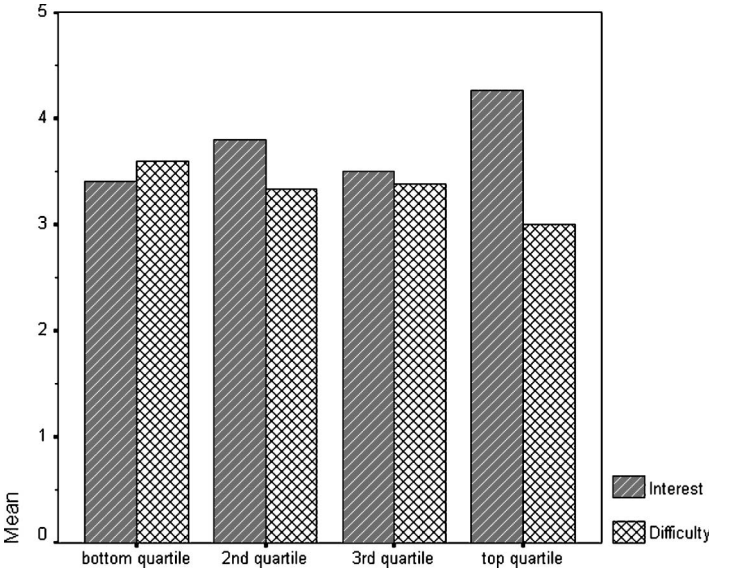

Figure 3. Mean ratings by quartile (1 = low, 5 = high interest/difficulty)

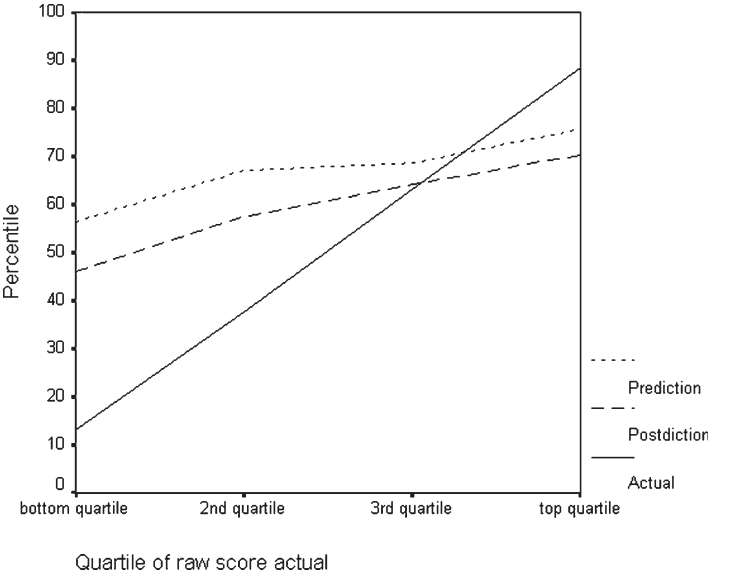

Figure 4. Relative percentile estimates by quartile

6. INSTRUCTOR CALIBRATION

Prior to administering the exam, we had believed that students would, on average, answer approximately 11 of the questions correctly, predicting that a large number would obtain perfect scores. Yet only a single student had a perfect score of 14, with no students scoring 13, four students scoring 12, and students scoring 8.39 on average. The exam questions avoided the algorithmic arcana that sometimes creep into data structures courses, asking instead about basic skills such as time-complexity analysis on iterative programs, code tracing, tree traversals, and algorithm identification. Perhaps we should not have been surprised with the student scores, given recent large-scale studies in which students underperformed their instructors' expectations (McCracken et al., 2001; Lister et al., 2004). Although poor student calibration can lead to metacognitive failures of learning and strategic choice, poor instructor calibration to student performance can lead to both instructor over-emphasis on material that is already mastered, or systemic blindness to student weaknesses that require remediation.

In order to determine if our colleagues were similarly miscalibrated to student performance, we carried out the following brief follow-up. After obtaining IRB approval, we presented instructors within our own departments with a copy of the exam (usually by email) and asked them the following single question: "Assuming that you are teaching a class that requires data structures as a prerequisite (e.g., AI, Algorithms) how many of the quiz questions from the attached quiz would you expect students to answer correctly on average? Please provide a real number between 0 and 14." We asked this of all instructors within our departments teaching either the

data structures course or one of the courses requiring data structures as a prerequisite.

Eleven instructors provided the following responses: 12, 10, 9.85, 12, 5.75, 12, 14, 10.5, 10.75, 10, 8, with a mean prediction of 10.44. As can be seen, all but two instructors overestimated the average student performance of 8.39. The average student prediction of 9.69 and postdiction of 8.99 were each better than the predictions of all instructors except one. In aggregate, student estimates were considerably more accurate than either aggregate or individual instructor estimates.

We conjecture that, while instructor illusions about student performance sustain hope, at the same time they hinder instructors from systematically determining the causes of student underperformance so as to provide remedial instructional interventions. But we leave it for subsequent studies to determine whether this is the case.

7. CONCLUSION

This study set out to answer a number of specific research questions about student knowledge and metaknowledge in computer science. We provide responses to each of these questions in light of the empirical study carried out. Additionally, we highlight the implications of this research for both the instructor and the Computer Science Education (CS Ed) researcher. We caution, however, that because this is a single, small-scale study, all conclusions must be viewed in the context of existing calibration research and as contingent on validation by future studies.

Do students have systematic misconceptions or lack of retention concerning the data structures material in subsequent courses? Test results revealed particular areas of weakness, including sorting and searching algorithms, hash tables, and tracing and analyzing recursive binary tree methods. However, it is unclear whether the weaknesses were due to insufficient learning in the first place or lack of retention after the data structures course ended. The fact that the students who had taken an algorithms course, which generally involves review of data structures, performed as poorly as those taking only data structures suggests that the issue is primarily one of recall; regardless, there is a considerable amount of data structures information that students do not have readily accessible. We state these conclusions cautiously, however, since students might perform differently over a larger, more comprehensive exam, in a non-test setting, or when writing code in a computer laboratory. Because performance on particular questions is likely influenced by the specifics of the instructional context, we would not claim that these areas of weakness generalize beyond our respective institutions. But at the risk of stating a truism, we observe that simply because a subject is ''covered'' in a course does not mean that students can access this information in the future when called upon to do so. Not only did the poor student performance on the exam ($M = 8.39$ out of 14) surprise us, it also surprised our colleagues who teach the data structures courses and the courses that depend upon it, with a mean instructor estimate of 10.44. Although a student's miscalibration affects only *that* student's metacognitive choices, the costs of an instructors' miscalibration are magnified across all students attending a course. The implications for both educators and researchers (even those doing informal research within their classrooms), are that researcher and instructor intuitions of student learning in the absence of substantiating empirical data can be significantly misleading.

Do student self-estimates correlate with performance, or do students systematically under- or over-estimate performance? Is calibration accuracy related to domain expertise? As a whole, students were quite well calibrated, with a correlation of 0.643 between postdicted and actual score, among the highest reported in the calibration literature we surveyed. Average postdiction error was also relatively low at 1.65 on an exam with 14 problems. Comparing these calibration results to those reported in the research literature, it is likely that the relatively high levels are due to a number of factors:

Domain expertise. The students sampled in these courses were students already achieving success in the major, and already possessing a large body of domain knowledge. Also, because the informed consent process occurred late in the academic term in all classes, some weaker students had already dropped the course.

Content of the exam. The contents of the introductory data structures course, at least in the United States, have been somewhat constant for a number of years (Tenenberg, 2003). This increases the likelihood that, regardless of instructor, students would have been exposed to similar material and asked to exercise similar skills in their data structures course. This might account for why even the prediction accuracy was quite high ($r = 0{:}42$, estimation error $= 2.3$), since students could develop a coherent predictive model of the material that they anticipated would be on the test. Even more importantly, much of the material is formal in nature, with precise criteria and processes for establishing right and wrong answers. That is, in reasoning about computer science knowledge in general, and data structures material in particular, students must have a complete chain of inference linking the given problem, their stored knowledge, and their conclusions. Thus, they might have a firmer grasp on when they are right and wrong than students working in disciplines (or on problems) where there are not such clear cut answers to problems. As Hans Bethe stated "In science, you know you know" (Associated Press, 2005).

Exam setting. In making predictions, students had the expectation that their self-assessment would be validated through an actual performance, one of the factors that Mabe and West (1982) found to be strongly correlated with calibration accuracy.

What this means for the instructor is that high levels of accurate self-knowledge cannot always be expected of students, since the above conditions are not always satisfied. Students working in unfamiliar knowledge domains (e.g. when learning domain knowledge specific to a software application), or domains that are less formal (e.g. Computer Ethics, Requirements Elicitation) might be less able to predict their performance. Open-ended and evaluative investigations might also diminish a student's calibration ability. And calibration accuracy might decrease in settings where students do not expect their knowledge claims to be validated. In these cases, it becomes much more important for students to have direct empirical validation of their performance so as to help uncover those areas where their intuitions about their abilities might mislead them. Lin and Zabrucky (1998) point out, however, that "self-generated feedback has a more positive impact on calibration than does other-provided feedback" (p. 384), suggesting that instructors should provide students with systematic practice with metacognitive techniques so that students will learn to regularly self-monitor their own learning.

Do students with higher scores have the illusion of knowing? Do those with lower scores suffer the dual burden of incompetence? Students in the top fifty percentile were well calibrated both before and after taking the test. It is possible that these students have deservedly high confidence in their general abilities, and so the specific questions that appeared on the exam mattered little for their self-estimates. Although students in the bottom fifty percentile had poorer predictive accuracy, their postdictive accuracy approached that of the higher performers. Not only did these lower performers not appear to suffer the dual burden of incompetence, they also demonstrated the capacity to calibrate from experience. For these

poorer performing students, prediction error might be more a result of having a faulty model of what comprises the materials in a data structures course rather than a failure to assess their own level of understanding.

Do most students rate themselves above average relative to their peers? Student estimates of performance relative to their peers were moderately calibrated. However, the lower two performance quartiles had considerably higher estimation errors than those in the top half, consistent with the well-known better-than-average effect. We conjecture a few reasons for this. First, in many courses, instructors do not report the distribution of grades across all students. Thus, self-beliefs about relative performance might be highly conjectural and unreliable. And second, students might reason that they are above average given their attainment thus far in the degree program, especially considering attrition out of the major that they have observed among many of their peers.

Perhaps most importantly, the better-than-average effect might have little influence on tactical metacognitive decision making, influencing instead larger strategical considerations, such as whether to continue in the current degree studies in Computer Science. It may be that for those students performing most poorly, both persistence in the major and preservation of self-image requires some amount of denial of relative ranking.

Does self-assessment of subject difficulty or interest correlate with either performance or accuracy? Neither student interest in the data structures material nor ratings of subject difficulty bore any relationship to either performance or knowledge calibration. This might be as a result of selection bias in the sample: students progressing this far in the major will likely be those who in general like the subject and do not find it overly difficult. The fact that students like a lesson, method, or course does not mean that they attained mastery, or that dislike means that they failed to do so. It suggests that neither instructors nor researchers should treat student reports about liking a teaching intervention or the difficulty of learning as a proxy (or dependent variable) for the amount of learning that actually occurred. Learning is much better assessed through direct evaluation of performance.

ACKNOWLEDGEMENTS

We are grateful to colleagues Donald Chinn and David Wolff for providing feedback on the quizzes and insight into the data structures courses that they teach at UWT and PLU; to Karen Furuya, Dan Bahrt, Kenneth Keeler, and Darrel Rohar for piloting versions of the quiz; to Bronwyn Pughe for editorial assistance; to the anonymous reviewers for their thoughtful comments 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.

NOTE

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

REFERENCES

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

Agnew, N., Ford, K., & Hayes, P. (1994). Expertise in Context: Personally Constructed, Socially Selected, Reality-Relevant? *International Journal of Expert Systems*, *7*(1), 65–88.

Associated Press (2005). Atom-bomb designer Hans Bethe dies. Retrieved March 7, 2005, from http://www.msnbc.msn.com/id/7118162/

Brown, A. (1987). Metacognition, Executive Control, Self-Regulation, and Other More Mysterious Mechanisms. In Weinert, F. and Kluwe, R., (Eds), *Metacognition, Motivation, and Understanding,* pages 65–116. Lawrence Erlbaum Associates, Inc.

College Entrance Examination Board (2003). *The College Board AP Advanced Placement Program Computer Science Course Description*. Retrieved July 20, 2003, from www.collegeboard.com.

Ehrlinger, J. & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology*, *84*(1).

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

Fitzgerald, T., Gruppen, L., White, C., & Davis, W. (1997). Medical Student Self-assessment Abilities: Accuracy and Calibration. In *Annual Meeting of the American Educational Research Association.*

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

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

Graduate Record Examination (2001). *Graduate Record Examinations Computer Science Test Practice Book*. Retrieved 20 July 2003 from www.gre.org.

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

Krueger, J. & Mueller, R.A. (2002). 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.

Kruger, J. (1999). Lake wobegone be gone! the ''below-average effect'' and the egocentric nature of comparative ability judgements. *Journal of Personality and Social Psychology*, *77*(2), 221–232.

Kruger, J. & Dunning, D. (1999). 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.

Lin, L.-M. & Zabrucky, K. (1998). Calibration of Comprehension: Research and Implications for Education and Instruction. *Contemporary Educational Psychology*, *23*, 345–391.

Lister, R., Adams, E.S., Fitzgerald, S., Fone, W., Hamer, J., Lindholm, M., McCartney, R., Moström, J. E., Sanders, K., Seppälä, O., Simon, B., & Thomas, L. (2004). A multi-national study of reading and tracing skills in novice programmers. *ACM SIGCSE Bulletin*, *36*(4), 119–150.

Mabe, III, P.A. & West, S.G. (1982). Validity of Self-Evaluation of Ability: A Review and Meta-Analysis. *Applied Psychology*, *67*(3), 280–296.

McCracken, M., Almstrum, V., Diaz, D., Guzdial, M., Hagan, D., Kolikant, Y. B.-D., Laxer, C., Thomas, L., Utting, I., & Wilusz, T. (2001). A multinational, multi-institutional study of assessment of programming skills of first year CS students. *ACM SIGCSE Bulletin*, *33*(4), 125–180.

Rammstedt, B. & Rammsayer, T. (2000). Sex differences in self-estimates of different aspects of intelligence. *Personality and Individual Differences*, *29*, 869–880.

Tenenberg, J. (2003). A framework approach to teaching Data Structures. In *34th SIGCSE Technical Symposium on Computer Science Education,* pp. 210–214.

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