Role of True-False Tests In Engineering Education

Laura W. Lackey, Ph.D., P.E.[1] and W. Jack Lackey, Ph.D.[2]

Abstract

True-false and open-ended test scores were compared for an undergraduate engineering course. There was good correlation between the scores indicating that the use of true-false test items was fair to students. Use of a mix of test item types is recommended since this permits testing over a greater fraction of the course content, encourages students to study, and is an efficient use of instructor time.

Introduction

As relatively new members of the teaching profession, we questioned whether true-false tests were appropriate for engineering students. We had seldom used this format; instead, relying on the more traditional problem-solving and essay type items. The desires to cover more of the course content per test and also to reduce grading time caused us to experiment by incorporating some true-false questions on each of two tests plus the final exam for an engineering course. This paper compares the students' scores individually and as a group on the true-false portion of the tests with the grades on the problem-solving/essay items. The overall goal was to determine if our tests and/or utilization of instructor time could be improved by use of some true-false questions.

This paper, particularly the literature review, is focused at assisting engineering faculty in preparing better exams. Testing is extremely important. It influences student study habits, job opportunities, and provides feedback to students and faculty. It has been estimated by Jacobs and Chase [1] that professors spend 20% of their time selecting, writing, administering, and scoring tests, yet have minimal formal education evaluating instruction. The results and analysis portions of the paper hopefully contribute new information that has broad applicability- particularly for engineering but also other disciplines.

Literature Review

No review of the effective development and implementation of tests should proceed without reference to the excellent book on the subject by Jacobs and Chase.[1] This book was written for university instructors and provided direct and concise answers to

[1] Laura W. Lackey, Assistant Professor, Environmental Engineering, 1400 Coleman Avenue, Mercer University, Macon, Georgia 31207-0001

[2] W. Jack Lackey, Professor, George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0405

many questions that had plagued us. Pertinent areas were: 1) appropriateness of various test items (multiple-choice vs true-false vs essay, etc.), 2) hints for writing better tests, 3) procedures for evaluating test reliability and validity, 4) administering tests, and 5) grading.

The book referenced above plus numerous papers report on studies conducted for the purpose of comparing different types of testing items. Comparison of multiple-choice with essay type questions has predominated.[2-7] Various forms of true-false tests have also been compared with multiple-choice, essay, and other types of questions.[3, 8-12] Most of the studies have been conducted for highschool or college courses on education, psychology, or health and medicine. Unfortunately, such studies have rarely been conducted in engineering education. Metrics of interest in such studies include: reliability (consistent measure of achievement),[10, 13] validity (true to course content),[10, 14] student preferences,[3, 14] ranking of students (grades), [14] time to complete by student, [9] student perception of both test worthiness and the instructor, [14, 15] and retention. [7, 16] The results of the comparisons are not always consistent. There are many variables, and most are difficult to control. The exact outcome of such analyses depend on skill and consistency in preparing the various test item types, level of student, study habits, etc. Nevertheless, considerable progress has been made, and most investigators would agree with the following summary.

The preferred test item depends on the objective of the test. For example, true-false and multiple-choice may be preferable to essay questions if the goal is to determine whether or not the students remember factual detail. If testing of higher levels of cognitive skills is desired, essay questions are preferable. We are using the commonly accepted categorization of cognitive skills measured by classroom tests provided by Bloom. [17] His six skills, varying from simple to complex, are: knowledge (recall of learned material), comprehension, application, analysis, synthesis, and evaluation. It should be noted that some investigators have concluded that skillfully-written true-false, or other test-type items normally labeled as objective items, can also be used to evaluate the higher cognitive skills. [1] Thus, generalizations regarding the merits of different type test questions is somewhat risky.

Despite the scatter in results, there is sufficient evidence to permit some sound conclusions. For example, true-false items are typically slightly less reliable than multiple-choice items. [1, 8] Objective items (true-false, multiple-choice, etc.) are more reliable than essay items since grading of essays is subjective. Jacobs and Chester [1] reviewed several sources for this claim. Essay grades depend on the order in which they occur in the stack. First papers that are read tend to get higher scores. Also, a paper read after an excellent paper tends to get a lower score. Poorly-constructed essay items are inefficient and may only require the student to simply reproduce the facts. Essay items should be reserved for demonstration of higher cognitive levels. Also, most agree that the disadvantages of objective test items compared to subjective items (essay) are balanced by the advantage that testing over a greater fraction of the course content is possible with objective items. In this regard, the true-false test is accepted by most as being superior. About 1.5 true-false items can be answered in the time required to answer one multiple-choice item. [1] It is also generally accepted that students study and

learn differently, depending on the anticipated test method. This has led most researchers to the important conclusion that the best test for most situations contains a mixture of test item types which covers as much as possible of the course content and which requires use of a range of cognitive skills. Certainly for engineering courses, problem-solving should also be a significant component of the test. Use of multiple test types encourages students to learn facts, theory, concepts, and how to apply them to new situations. [1] In other words, the goal is to write tests that measure a variety of skills, especially at the higher cognitive levels. A varied test format is thought to improve retention, motivate students to study, and enhance the students' perception of both the test and the instructor.

The emphasis of the current paper is true-false items. Jacobs and Chase [1] and Barker and Ebel [18] give the following suggestions for writing good true-false tests:

- Avoid use of specific determiners such as "all," "always," etc. since they indicate that the item is probably "false."
- Avoid use of "sometimes," "usually," "typically," etc. since they indicate that the item is probably "true."
- Avoid use of indefinite words or phrases denoting degree or amount.
- True-false items should be positive, declarative sentences, stated as simply as possible.
- Compound sentences can be used effectively. For example, the first part of the sentence (stem) may describe a true condition or setting which is followed by a clause which the student must decide is true or false.
- The condition statement may be a short paragraph rather than a clause, and multiple true-false questions can be asked in the same setting.

Method

Participants

Participants in the study were 38 mechanical engineering students enrolled in an undergraduate materials engineering course. The course was required and it met three times a week for a semester. The students were juniors and seniors and were not aware of the study.

Procedure

Two 50-minute tests and a 2.8-hour comprehensive final exam were administered. The first test consisted of 40 true-false questions that counted for 60% of the total score. The remainder of Test No. 1 included mostly free response items (like essays) which required answers varying from a few sentences to half a page. Two items involved sketches, and another directed them to recall an equation and to describe each term. The true-false items were not formatted to cover the exact topics as the essay items in this, nor subsequent tests. The second test had 18 true-false items that counted for 36% of the score. The remainder of Test No. 2 consisted of 15 items that required answers of only one or a few words. A third test was administered but was not analyzed since it did not contain any true-false items. The final exam consisted of 60 true-false items worth 30% of the score. The remainder was comprised of a mix of items similar in type to the non true-false part of Test No. 1. One student did not take Test No. 2 and the final exam because of sickness, and two others were graduating and took a different final exam. These three students were not included in the study.

All tests were announced two or more weeks in advance. The students were not aware that true-false items would be included on Test No. 1, but did expect Test No. 2 and the final exam to contain true-false as well as other item types.

Scoring

The total possible score on each test was 100. All students answered each of the true-false items and there was no additional penalty for erroneous responses to true-false items. Partial credit was given for the non true-false items. The students' names were hidden during scoring, the non true-false items were scored one question at a time, and papers were shuffled after reading each non true-false item.

Analysis

The primary interest was to determine whether the students scored similarly on the true-false portion and remaining portion of each test. To this end, least squares linear regression was performed for each test and correlation coefficients were calculated using the statistical package STATGRAPHICS.

Results and Discussion

The scores for the true-false and non true-false portions of the two tests and final exam are given in Table 1. The correlation coefficients obtained when comparing the true-false and non true-false scores for the two tests and final exam were 0.57, 0.43, and 0.56, respectively. The Spearman-Brown formula [1] was used to calculate reliability. The values were 0.73, 0.60, and 0.72 for the two tests and final exam, respectively. These values indicate that there was reasonable agreement between the scores for the

true-false and the remaining portion of each of the three tests. The relationship between scores for Test No. 1 is shown graphically in Figure 1. It is apparent that as the non true-false score increases, the true-false score typically increases even though there is appreciable scatter in the data. This same trend existed for the other two tests.

Two lines are shown in Figure 1. One is the 45° line going through the points 50,50 and 100,100. The point representing a student who received the same score on the true-false and remaining portion of the test would fall on this line. More points are located above this line than below. That is, scores for the true-false items (average = 83.3) were higher than scores for the remainder of Test No. 1 (average = 80.7). Several factors could have caused the true-false scores to be higher. One such factor is guessing. For example, if a student knows 80% of the material covered on the test, the expected score for the non true-false segment would be 80. This same student would typically not know 20% of the true-false items. Assuming he/she would guess correctly half of the time, the expected score on the true-false portion of the test would be 90. Similar reasoning led to the upper line in the Figure which is labeled “Expected Relationship.” Note that the point just discussed (80,90) is located on this line. More data points are located below this line than above, indicating that the true-false test items may have been slightly more difficult than the remainder of the test. In an earlier version of Figure 1, and the following two figures, a number identifying the student was placed adjacent to each data point in an effort to determine any tendency for a student to consistently perform better on a particular item type. No such tendency was observed.

Figures 2 and 3 show similar plots for Test No. 2 and the final exam. All of the observations made above for Test No. 1 are also applicable to Test No. 2. That is, the same trends and tendencies, or lack thereof, were observed. This is not true for the final exam as is apparent in Figure 3. The data points were located below the Expected Relationship line, perhaps indicating that the true-false part of the exam was more difficult than the remainder of the exam.

The final exam was sufficiently long that the split-half procedure [1] for determining reliability was employed. That is, a score was calculated for each student for the odd-numbered true-false items and for the even-numbered ones. This yielded a Spearman-Brown value for reliability of 0.55. When this procedure was repeated for the non true-false items, a reliability value of 0.79 was obtained. This indicates that the non true-false part of the final exam was more reliable. This was expected since it was longer, i.e. counted for 70% of the overall score.

Grades are very important to students. While we did not rigorously evaluate whether the final ranking of students would have changed if true-false items had not been used, the reasonably high correlation coefficients between the true-false and non true-false items make it unlikely that student rankings would have been significantly influenced.

Table 1. True-False and Non True-False Scores

	TEST NO. 1		TEST NO. 2		FINAL EXAM	
	True-False Score	Non True-False Score	True-False Score	Non True-False Score	True-False Score	Non True-False Score
	90	72.5	67	52	83	81
	82.5	76.2	78	59	77	61
	79.2	96.2	61	44	85	78
	95	87.5	78	72	83	92
	82.5	88.7	67	56	80	84
	92.5	88.8	83	61	90	90
	90	87.5	61	59	73	85
	60	57.5	50	55	72	59
	87.5	63.8	78	58	63	58
	88.3	100	61	64	82	88
	80	70	61	58	73	53
	57.5	36.2	56	23	58	39
	85	72.5	83	50	80	78
	93.3	97.5	44	41	70	60
	92.5	96.2	89	67	78	87
	90	100	78	73	90	92
	85	87.5	50	41	73	67
	77.5	86.2	44	50	65	72
	77.5	96.2	78	64	78	74
	87.5	98.8	94	56	72	84
	80	95	78	77	88	81
	82.5	68.8	67	62	83	68
	82.5	68.8	94	53	70	85
	82.5	78.8	89	41	72	69
	82.5	75	67	39	73	65
	80	75	83	52	75	82
	82.5	73.8	56	28	63	73
	92.5	81.2	67	66	82	76
	85	55	67	47	80	79
	87.5	83.8	83	72	80	89
	87.5	83.8	83	67	80	85
	82.5	81.2	56	34	67	75
	87.5	78.8	78	42	70	92
	77.5	83.8	67	23	65	64
	87.5	96.2	56	38	75	65
	85	85	67	52	83	67
	82.5	68.8	67	52	77	90
	65.8	73.8	78	39	78	92
Average	83.3	80.7	70.1	52.3	75.9	75.8
Std. Dev.	8.1	13.9	13.3	13.6	7.6	12.7

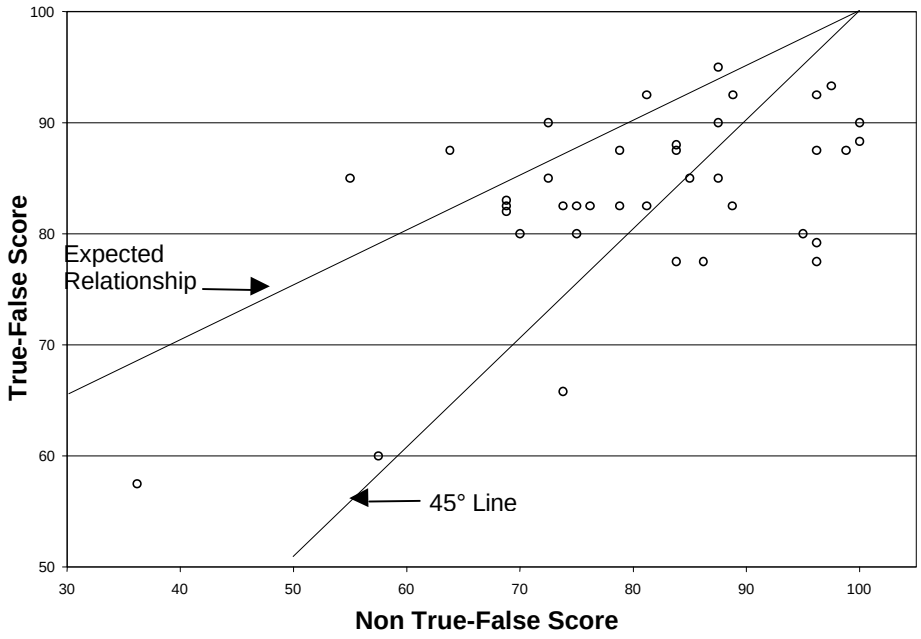

Figure 1. True-False Score Correlates Well with Non True-False Score for Test No. 1.

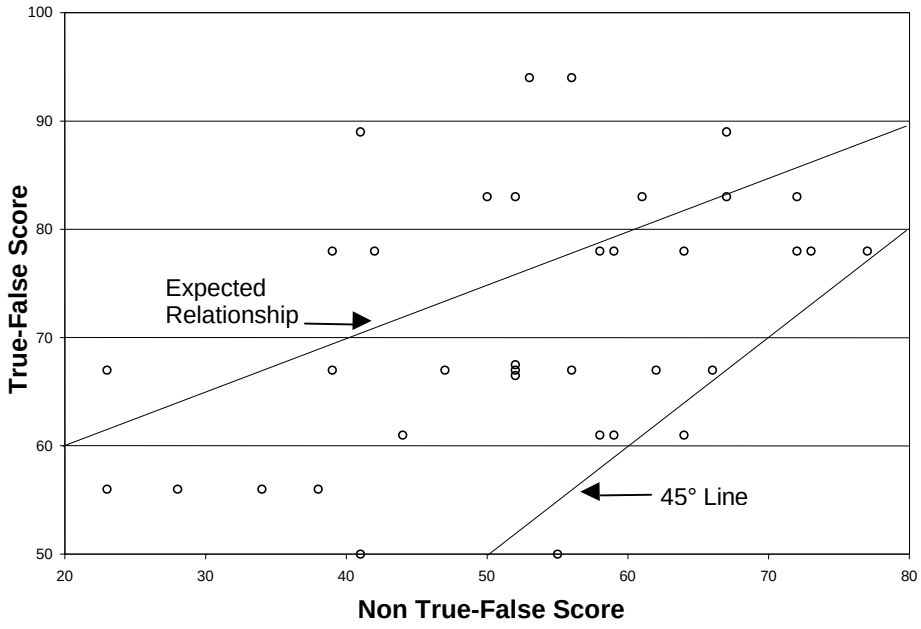

Figure 2. Comparison of Scores of Test No. 2.

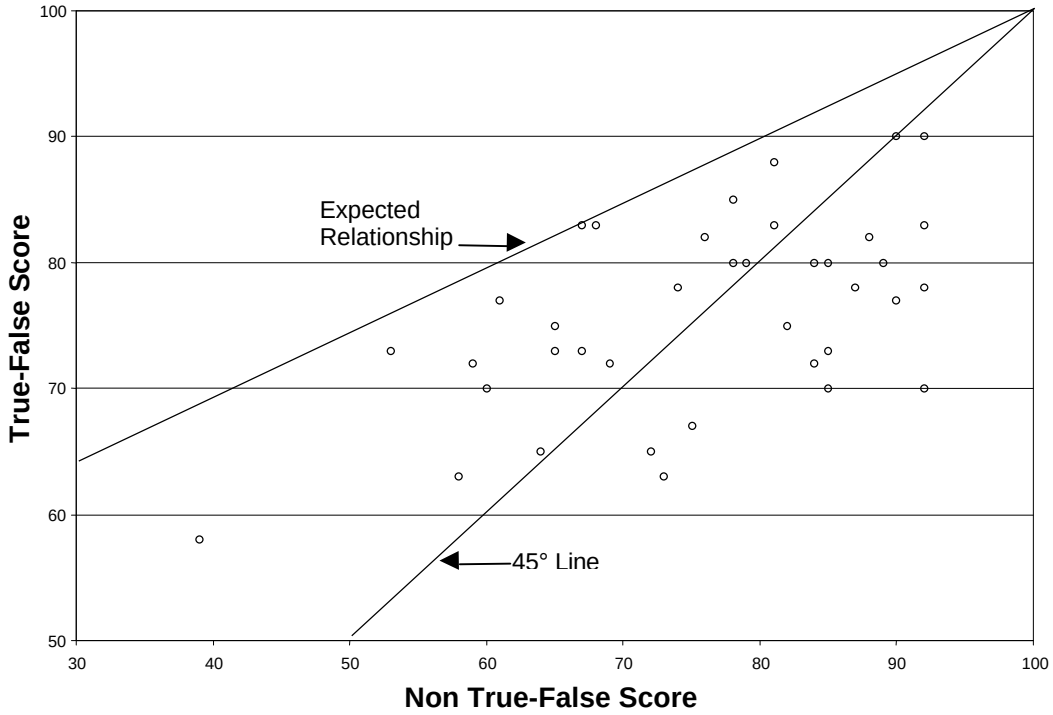

Figure 3. Comparison of Scores for the Final Exam.

Summary and Conclusions

The most important outcome of this study is that we, and hopefully the reader, now feel confident that the use of true-false items on tests is not only appropriate but is beneficial. Their use clearly permits testing over a larger fraction of the course content and appear to be fair to students. We and prior researchers suggest use of multiple test item types on any given test, with attention given to properly matching course topics with item types. We agree with Cirn [14] that it would be good for faculty to occasionally perform the analyses presented in this paper. If reliability values are low or the scores derived from true-false or other objective item types differ significantly from scores for more open-ended type items, then the instructor should consider improving his or her test-writing skills.

References

1. Jacobs, Lucy Cheser and Clinton I. Chase (1992) *Developing and Using Tests Effectively*, Jossey-Bass Publishers, San Francisco, California.

2. Gay, Lorraine R. (1980) "The Comparative Effects of Multiple-Choice versus Short-Answer Tests on Retention," *Journal of Educational Measurement*, 17(1) 45-50.

3. Anderson, Paul S. (1998) "An Educology of Testing: American Student Attitudes about Test Formats, with Special Reference to the MDT Multi-Digit Testing Technique," *International Journal of Educology*, 2(2) 143-184.

4. Anbar, Michael (1991) "Comparing Assessments of Students' Knowledge by Computerized Open-Ended and Multiple-Choice Tests," *Academic Medicine,* 66(7) 420-422.

5. Bennett, Randy Elliot, Donald A. Rock, and Minhwei Wang (1991) "Equivalence of Free-Response and Multiple-Choice Items," *Journal of Educational Measurement*, 28(1) 77-92.

6. Hancock, Gregory R. (1994) "Cognitive Complexity and the Comparability of Multiple-Choice and Constructed-Response Test Formats," *Journal of Experimental Education,* 62(2) 143-157.

7. Zopp, William B. (1998) "A Study of Comparing Testing Styles and Retention of Material," M.A. Thesis in Education, Salem-Teikyo University, Salem, West Virginia.
8. Ebel, R. L. (1975) "Can Teachers Write Good True-False Test Items?" *Journal of Educational Measurement*, 12, 31-36.

9. Green, Kathy (1979) "Multiple Choice and True-False: Reliability and Validity Compared," *Journal of Experimental Education,* 48(1) 42-44.

10. Sax, Gilbert and Pauline B. Reiter (1980) "Reliability and Validity of Two-Option Multiple-Choice and Comparably Written True-False Items," Technical Report TM 830-681, Available from Department of Education, National Institute of Education, Educational Resources Information Center, Document Reproduction Services as ED 236-177.

11. Frisbie, David A. and Daryl C. Sweeney (1982) "The Relative Merits of Multiple True-False Achievement Tests," *Journal of Educational Measurement,* 19(1) 29-35.

12. Maihoff, N. A. and William A. Mehrens (1985) "A Comparison of Alternate-Choice and True-False Item Forms used in Classroom Examinations," Presented at the Annual Researchers Meeting of the National Council on Measurement in Evaluation, Chicago, Illinois, April 1-3, Technical Report TM 850-475, Available from Department of Education, National Institute of Education, Educational Resources Information Center, Document Reproduction Services as ED 269-411.

13. Oosterhot, Albert C. and Pamela K. Coats (1981) “Comparison of Difficulties and Reliabilities of Math-Completion and Multiple-Choice Item Formats,” Presented at the Annual Meeting of the 65th American Educational Research Association, Los Angeles, California, April 13-17, Technical Report TM 810-723, Available from U.S. Department of Education, National Institute of Education, Educational Resources Information Center, Document Reproduction Services as ED 208-028.

14. Cirn, John T. (1986) “True-False versus Short Answer Questions,” *College Teaching* 34(1) 34-37.

15. Handleman, Chester (1974) “The Relationship Between Objective versus Subjective Classroom Tests and Student Evaluations of Their Instructors,” Ed.D. Practicum in Education, Nova Southeastern University, Fort Lauderdale, Florida, Available from U.S. Department of Education, National Institute of Education, Educational Resources Information Center, Document Reproduction Services as ED 110-144.

16. Duchastel, Philippe C. (1980) “Retention of Prose Following Testing with Different Types of Tests,” Paper presented at the Annual Meeting of the American Educational Research Association, Boston, Massachusetts, April 7-11, Available from U.S. Department of Education, National Institute of Education, Educational Resources Information Center, Document Reproduction Services as ED 182-727.

17. Bloom, B. S. (editor) (1956) *Taxonomy of Educational Objectives*, *Volume 1: Cognitive Domain*, McKay, New York.

18. Barker, Douglas and Robert L. Ebel (1982) “A Comparison of Difficulty and Discrimination Values of Selected True-False Item Types,” *Contemporary Educational Psychology*, 7(1) 35-40.

Laura W. Lackey, Ph.D., P.E.

Dr. Lackey has two years of experience as an assistant professor of Environmental Engineering at Mercer University. She has six years of industrial experience at the Tennessee Valley Authority as an Environmental/Chemical Engineer where she conducted both basic and applied research with emphasis on the mitigation of organic wastes through bioremediation. She earned a BS, MS, and Ph.D. in Chemical Engineering from the University of Tennessee. The terminal degree was awarded in 1992. She is a registered professional engineer in Alabama.

W. Jack Lackey, Ph.D.

Dr. Lackey received B.S. degrees in Ceramic Engineering and Metallurgical Engineering from North Carolina State University in 1961. He received a Master of Science degree and Ph.D. in Ceramic Engineering from North Carolina State University in 1963 and 1970, respectively. He conducted basic and applied research on nuclear fuel fabrication, nuclear waste disposal, and processing of ceramic coatings and composites at Battelle Northwest Laboratory and the Oak Ridge National Laboratory. From 1986-1997 while employed at the Georgia Tech Research Institute, he performed research on ceramic coatings and composites, advised graduate students in Materials Science and Engineering and Chemical Engineering, and taught undergraduate and graduate courses on mechanical behavior of materials and ceramic composite processing. In 1997 he joined the faculty of the George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, as a Professor.

Since 1997, Dr. Lackey has developed and taught an undergraduate course on materials selection and failure analysis and two graduate courses emphasizing processing of advanced ceramic coatings and composites and the interrelationships between processing, microstructure, and material properties. He has also taught an undergraduate course in materials science and engineering and a graduate course on nuclear materials.

Dr. Lackey currently advises eight graduate students. Their research areas are: 1) laser chemical vapor deposition rapid prototyping of electronic devices, carbon nanotubes, and structural nanolaminates, 2) processing of fiber-reinforced composites possessing a laminated matrix for enhancing fracture toughness, and 3) development of an improved process for carbon-coating of mechanical heart valves. He has published 90 refereed papers and has 15 patents.

He is the proud father of co-author Laura W. Lackey.