A Methodological Review of Computer Science Education Research

Justus Randolph
University of Joensuu, Joensuu, Finland
justus@randolph.name

George Julnes
Utah State University, Logan, Utah, USA
george.julnes@usu.edu

Erkki Sutinen
University of Joensuu, Joensuu, Finland
erkki.sutinen@cs.joensuu.fi

Steve Lehman
Utah State University, Logan, Utah, USA
slehman@cc.usu.edu

Executive Summary

Methodological reviews have been used successfully to identify research trends and improve research practice in a variety of academic fields. Although there have been three methodological reviews of the emerging field of computer science education research, they lacked reliability or generalizability. Therefore, because of the capacity for a methodological review to improve practice in computer science education and because the previous methodological reviews were lacking, a large scale, reliable, and generalizable methodological review of the recent research on computer science education is reported here. Our overall research question, which has nine subquestions, involved the methodological properties of research reported in articles in major computer science education research forums from the years 2000-2005. The purpose of this methodological review is to provide a methodologically rigorous basis on which to make recommendations for the improvement of computer science education research and to promote informed dialogue about its practice.

A proportional stratified random sample of 352 articles was taken from a population of 1306 computer science education articles published from 2000 to 2005. The 352 articles were coded in terms of their general characteristics, report elements, research methodology, research design, independent, dependent, and mediating/moderating variables examined, and statistical practices. A second rater coded a reliability sub-sample of 53 articles. Based on the results of this review, recommendations for improving computer science education research are given.

Keywords: Computer science education, computer science, research, methodological review, research methods, research practice, meta-analysis

Editor: Zlatko Kovacic

Introduction

One of the most influential books on computer science education research (Fincher & Petre, 2004) begins with the following statement: "Computer science education research is an emergent area and is still giving rise to a literature" (p. 1). Top computer science education researchers like Mark Guzdial and Vicki Almstrum argue that the interdisciplinary gap between computer science education and educational research proper, including methods developed in the broader field of behavioral research, must be overcome before computer science education research can be considered to be a field which has emerged (Almstrum, Hazzan, Guzdial, & Petre, 2005). (In this methodological review, we use the term *behavioral research* as a synonym for what Guzdzial, in Almstrum et al. (2005, p. 192), calls "education, cognitive science, and learning sciences research.") Addressing this lack of connection with behavioral research, Guzdial, in Almstrum and colleagues (2005) wrote,

> The real challenge in computer education is to avoid the temptation to re-invent the wheel. Computers are a revolutionary human invention, so we might think that teaching and learning about computers requires a new kind of education. That's completely false: The basic mechanism of human learning hasn't changed in the last 50 years. Too much of the research in computing education ignores the hundreds of years of education, cognitive science, and learning sciences research that have gone before us. (pp. 191-192)

One way to bridge the gap between computer science education research and behavioral research proper is to review current methodological practices in computer science education and to compare those practices with existing knowledge about best practices from the field of behavioral research. In this article, we do just that; we review the current methodological practices in computer science education and present recommendations, from a behavioral science standpoint, on how computer science education research might be improved. It is our hope that our results and recommendations will improve practice and inform policy about and, ultimately, help computer science education research transition from an emerging research area to an area that has already emerged. (Here we use Denning et al.'s (1989, p. 12) definition of the discipline of computing: "the systematic study of algorithmic processes that describe and transform information: their theory, analysis, design, efficiency, implementation, and application.")

Several groups stand to gain from this review. The creators of computer science education research will benefit from knowledge of how their peers tend to conduct research—such as what variables they investigate, how they tend to measure those variables, and how they analyze and present their data—and from getting suggestions on how to improve their own research. The consumers of computer science education research, such as funders, practitioners, and educational administrators, will become more aware of the strengths and weakness of the current computer science education research and can use that knowledge to make decisions about policy or practice. Finally, the gatekeepers of computer science education, such as the funders, editors, and reviewers of computer science education research, are especially important stakeholders in this review because they set the standard for what types of research activities are acceptable and what types of reports merit publication.

The next section begins with a discussion of three pre-existing reviews of the computer science education research and a rationale for the need for the current research. In the remaining sections of this paper, after a short discussion of biases, the reader will find a description of the methods used, including a description of the coding book development, the sampling strategy and sampling frame, interrater training procedures, and data analyses. In the results section, the reader can find a wide variety of descriptive statistics on computer science education research, from the most prolific authors to the types of statistical analyses typically conducted. In the discussion section,

we revisit our research questions. This article ends with recommendations for improving research practice and a short conclusion.

Literature Review

To inform the current research, we looked for other previous reviews of the computer science education research. To identify those reviews, we conducted repeated electronic searches of major academic databases, including the ACM digital library; searched the table of contents of major computer science education publications, like the SIGCSE Bulletin; and did a branching search of references of relevant articles. Three previous methodological reviews of the computer science education research involving either secondary or postsecondary participants was found: Randolph, Bednarik, and Myller (2005), Randolph (in press), and Valentine (2004). A short summary of each of these reviews is provided below.

A Review of K-12 Computer Science Education Program Evaluations

Randolph (in press) conducted a methodological review and meta-analysis of the program evaluation reports in computer science education. (Throughout this methodological review, because of the difficulties of making an external decision about what is *research* and what is *evaluation*, we operationalize an *evaluation report* as a document that the author called an *evaluation, evaluation report,* or *a program evaluation report*.) After an electronic and hand search of major academic databases, Randolph (in press) found 29 program evaluation reports of K-12 computer science education programs.

The major findings of Randolph (in press) are summarized below:

- Most of the programs that were evaluated offered direct computer science instruction to general education, high school students in North America.
- In order of decreasing frequency, evaluators examined stakeholder attitudes, program enrollment, academic achievement in core courses, and achievement in computer science.
- The most frequently used measures were, in decreasing order of frequency, questionnaires, existing sources of data, standardized tests, and teacher-made or researcher-made tests. Only one computer science achievement instrument, which is no longer available, had reliability or validity estimates.
- The pretest-posttest design with a control group and the one-group posttest-only design were the most frequently used research designs.
- No interaction between type of program and computer science achievement improvement was found.

In terms of the link between program evaluation and computer science education, the fact that there were so few program evaluations being done, that so few of them (i.e., only eight) went beyond simple program description and student attitudes, that only one used an instrument with information about measurement reliability and validity, and that one-group posttest-only designs were so frequently used indicate that the past K-12 computer science education program evaluations have had many deficiencies. As the next review indicates, the deficiencies are not solely found in K-12 computer science education program evaluations; there are also several deficiencies in computer science education research in higher education as well.

A Methodological Review of Selected Articles in SIGCSE Technical Symposium Proceedings

Valentine (2004) critically analyzed over 20 years of computer science education conference proceedings that dealt with first-year university computer science instruction. In that review, Valentine categorized 444 articles into six categories. The major finding from his review was that only 21% of papers in the last 20 years of proceedings were categorized as *experimental*, which was operationalized as the author of the paper making "any attempt at assessing the 'treatment' with some scientific analysis" (p. 256). Some of Valentine's other findings are listed below:

- The proportion of experimental articles had been increasing since the mid-90s.
- The proportion of what he calls *Marco Polo—I went there and I saw this*—types of papers had been declining linearly since 1984.
- The overall number of papers being presented in the SIGCSE forum had been steadily increasing since 1984. (As cited in Randolph, Bednarik, & Myller, 2005, p. 104)

Valentine concluded that the challenge is to increase the number of experimental investigations in computer science education research and decrease the number of "I went there and saw that," self-promotion, or descriptions-of-tools types of articles. The reliability of Valentine's findings, however, is questionable; Valentine was the single coder and reported no estimates of interrater agreement.

A Methodological Review of the Papers Published in Koli Calling Conference Proceedings

Randolph, Bednarik, and Myller (2005) conducted a critical, methodological review of all of the full-papers in the proceedings of the *Koli Calling: Finnish/Baltic Sea Conference on Computer Science Education* (hereafter *Koli Proceedings)* from 2001 to 2004. Each paper was analyzed in terms of (a) methodological characteristics, (b) section proportions (i.e., the proportion of literature review, methods, and program description sections), (c) report structure, and (d) region of origin. Based on an analysis of all of the full-papers published in four years of Koli Proceedings, their findings were that

- The most frequently published type of paper in the Koli Proceedings was the program (project) description.
- Of the empirical articles reporting research that involved human participants, survey research and quasi-experimental methods were the most common.
- The structure of the empirical papers that reported research involving human participants deviated sharply from structures that are expected in behavioral science papers. For example, only 50% of papers that reported research on human participants had literature reviews; only 17% had explicitly stated research questions.
- Most of the text in empirical papers was devoted to describing the evaluation of the program; very little was devoted to literature reviews.
- The Koli Calling proceedings represented mainly the work of Nordic/Baltic, especially Finnish, computer science education researchers.
- An additional finding was that no article reported information on the reliability or validity of the measures that were used.

Both the Valentine (2004) and Randolph, Bednarik, and Myller (2005) reviews converged on the finding that few computer science education research articles went beyond describing program activities. In the rare cases when impact analysis was done, it was usually done using anecdotal evidence or with weak research designs.

The Scope and Quality of the Previous Methodological Reviews of Computer Science Education Research

Although there have been three methodological reviews of research on computer science education, those reviews were limited either in their breadth, depth, or reliability. The three previous methodological reviews of computer science education research (Randolph, in press; Randolph, Bednarik, & Myller, 2005; Valentine, 2004) cover only a very small part of the population of mainstream, recent computer science education research. Additionally, the review that is most representative of the population of computer science education research articles (i.e., Valentine) has limited generalizability and reliability: Valentine reviewed a large number of articles, but he sampled them from only one forum for publishing computer science education research and excluded articles that did not deal with first-year computer science courses. Also, Valentine coded all of the articles himself without any measure of interrater agreement and he only coded one variable for each article. He simply classified the articles into one of six categories: *Marco Polo, Tools, Experimental, Nifty, Philosophy*, and *John Henry*. The *experimental* category—operationalized as "any attempt at assessing the 'treatment' with some scientific analysis" (Valentine, 2004, p. 256)—is so broad that it is not useful as a basis for recommending improvements in practice.

Purpose and Research Questions

Because the past methodological reviews of computer science education research had limitations either in terms of their generalizability or reliability, we conducted a replicable, reliable, methodological review of a representative sample of the research published in the major computer science education forums over the last 6 years. This methodological review (a) provides significantly more-representative coverage of the field of computer science education than any of the previous reviews, (b) covers articles with more analytical depth (with a more-refined coding sheet) than any of the previous reviews, and (c) with a greater amount of reliability and replicability than any of the other previous reviews. In short, this methodological review simultaneously extends the breadth, depth, and reliability of the previous reviews.

The purpose of this methodological review was to have a methodologically rigorous basis on which to make recommendations for the improvement of computer science education research and to promote informed dialogue about its practice. If our recommendations are heeded and dialogue increases, computer science education is expected to improve and, consequently, help meet the social and economic needs of a technologically oriented future.

We answered the primary research question: What are the methodological properties of research reported in articles in major computer science education research forums from the years 2000-2005? The primary research question was broken down into several sub-questions, which are listed below:

1. What was the proportion of articles that did not report research on human participants?
2. Of the articles that did not report research on human participants, what types of articles were being published and in what proportions?

3. Of the articles that did report research on human participants, what proportion provided only anecdotal evidence for their claims?
4. Of the articles that did report research on human participants, what types of methods were used and in what proportions?
5. Of the articles that did report research on human participants, what measures were used, in what proportions, and was psychometric information reported?
6. Of the articles that did report research on human participants, what were the types of independent, dependent, mediating, and moderating factors that were examined and in what proportions?
7. Of the articles that used experimental/quasi-experimental methods, what types of designs were used and in what proportions? Also, were participants randomly assigned or selected?
8. Of the articles that reported quantitative results, what kinds of statistical practices were used and in what proportions?
9. Of the articles that did report research on human participants, what were the characteristics of the articles' structures?

Biases

The backgrounds of the principal investigator (the first author) and the second and fourth authors are in behavioral science research (particularly quantitative education research and program evaluation). The third author has a background in computer science. Primarily, we brought the biases of quantitatively trained behavioral scientists into this investigation. It is our belief that when one does education-related research on human participants the conventions, standards, and practices of behavioral research should apply. Nevertheless, we realize that computer science education and computer science education research is a maturing, multidisciplinary field, and we acknowledge that the behavioral science perspective is just one of many valid perspectives that one can take in analyzing computer science education research.

Method

Neuendorf's (2002) *Integrative Model of Content Analysis* was used as the model for carrying out the methodological review. Neuendorf's model consists of the following steps: (a) developing a theory and rationale, (b) conceptualizing variables, (c) operationalizing measures, (d) developing a coding form and coding book, (e) sampling, (f) training and determining pilot reliabilities, (g) coding, (h) calculating final reliabilities, and (i) analyzing and reporting data. In the following subsections, we describe how we conducted each of the steps of Neuendorf's model. Because the rationale (the first step in Neuendorf's model) was described earlier, we do not discuss it below.

Conceptualizing Variables, Operationalizing Measures, and Developing a Coding Form and Coding Book

Because this methodological review was the sixth in a series of methodological reviews we had conducted (see Randolph, 2007b; Randolph, in press; Randolph, Bednarik, & Myller, 2005; Randolph, Bednarik, Silander, et al., 2005; Randolph & Hartikainen, 2005; and Randolph, Hartikainen, & Kähkönen, 2004), most of the variables had already been conceptualized, measures had been operationalized, and coding forms and coding books had been created in previous reviews.

A list of the articles that were sampled is included in Appendix A of Randolph (2007a). The coding form and coding book that we used for this methodological review are included as Appendices B and C of Randolph (2007a), respectively.

Sampling

A proportional stratified random sample of 352 articles, published between the years 2000 and 2005, was drawn, without replacement, from eight major peer-reviewed computer science education publications. (That sample size, 352, out of a finite population of 1,306 was determined a priori, through the *Sample Planning Wizard* (2005) and confirmed through resampling.) The sample was stratified according to year and source of publication. Table 1 shows the sample size and populations by year and forum. The 352 articles that were included in this sample are listed in Appendix A of Randolph (2007a).

Table 1. Sample and Population Sizes

Year/forum	2000	2001	2002	2003	2004	2005	Total
Bulletin	8/31	6/21	6/21	11/40	10/36	10/38	51/187
CSE	5/17	5/17	5/17	5/17	5/17	4/ 15	29/100
JCSE	0/0	0/2	2/7	1/ 5	0/2	0/2	3/18
KOLI	0/0	4/14	3/10	3/13	6/21	7/25	23/83
SIGCSE	21/78	21/78	20/74	20/75	25/92	28/104	135/501
ITICSE	12/45	12/44	11/42	11/41	12/46	18/68	76/286
ICER	0/0	0/0	0/0	0/0	0/0	4/16	4/16
ACE	0/0	0/0	0/0	9/34	13/48	/933	31/115
Total	**46/171**	**48/176**	**47/171**	**60/225**	**71/262**	**80/301**	**352/1306**

Note. The first number in each cell is the sample size. The second number is the population size.

The population was operationalized in such a way that it was a construct of what typically is accepted as mainstream computer science education research. We did not include forums that were not devoted primarily to computer science *education,* such as the *Journal of Information Systems*. The population did not include the marginal, gray areas of the literature, such as unpublished reports, program evaluation reports, or other nonpeer-reviewed publications, because we were not interested in the research practices reported in the entirety of computer science education research. Rather, we were interested in research practices reported in current, peer-reviewed, mainstream computer science education research forums. We operationalized these as the June and December issues of *SIGCSE Bulletin* (hereafter *Bulletin*), a computer science education journal; *Computer Science Education* (hereafter *CSE*), a computer science education research journal; the *Journal of Computer Science Education Online*, (hereafter *JCSE*), a little-known computer science education journal; the *Proceedings of the SIGCSE Technical Symposium* (hereafter *SIGCSE*); The *Proceedings of the Innovation and Technology in Computer Science Education Conference* (hereafter *ITiCSE*); the *Koli Calling: Finnish/Baltic Sea Conference on Computer Science Education* (hereafter *Koli*), the *Proceedings of the Australasian Computing Education Conference* (hereafter ACE), and the *International Computer Science Education Research Workshop* (hereafter *ICER*). (The fall and spring issues of *Bulletin* are the *SIGCSE* and *ITiCSE* proceedings.) In hindsight, we acknowledge that we probably should have included the *Journal of Information Technology Education* in the sample, but it was unknown to us at the time. We included "full papers," but excluded poster summaries, demo summaries, editorials, conference reviews, book reviews, forewords, introductions, and prologues in the sampling frame.

In general, nonpeer-reviewed articles or poster-summary papers (i.e., papers two or fewer pages in length) were not included in the sampling frame. In *Bulletin*, only the peer-reviewed articles were included; featured columns, invited columns, and working group reports were excluded in the sampling frame of Table 1. In *CSE* and *JCSE*, editorials and introductions were excluded. In the *SIGCSE*, *ITICSE*, *ACE*, and *ICER* forums, only full peer-reviewed papers at least three pages in length were included; panel sessions and short papers (i.e., papers two pages or less in length) were excluded. In *Koli,* research and discussion papers were included; demo and poster papers were excluded.

Training and Determining Pilot Reliabilities

An interrater reliability reviewer, who had participated in previous methodological reviews, was trained in the coding book and coding sheet, which are included as Appendices B and C of Randolph (2007a). Initially the interrater reliability reviewer and the first author read through the coding book and coding sheet together and discussed any questions about the coding book or coding sheet. When inconsistencies or ambiguities in the coding book or coding sheet were found in the initial training session, the coding book or coding sheet was modified to remedy those inconsistencies or ambiguities. Then, the interrater reliability reviewer was given a revised version of the coding book and coding sheet and was asked to independently code a purposive pilot sample of 10 computer science education research articles, which were not the same articles that were included in the final reliability subsample. The purposive sample consisted of articles that the first author deemed to be representative of the different types of research methods that were to be measured, articles that were anecdotal only, and articles that did not deal with human participants. The primary coder, the first author, also coded those 10 articles. After both of coders had coded the 10 articles they came together to compare the codes and to discuss the inconsistencies or unclear directions in the coding book and coding sheet. When the coders had disagreements about article codes, the coders would try to determine the cause of the disagreement and the first author would modify the coding book if it were the cause of the disagreement. After pilot testing and subsequent improvement of the coding book and the coding, the final reliability subsample was coded (see the section entitled *Calculating Final Reliabilities*).

Coding

Appendices B and C of Randolph (2007a), which are the coding sheet and coding book, provide detailed information on the coding variables, their origin, and the coding procedure. Because the complete coding sheet and coding book are available as appendices in Randolph (2007a), we do not report on them here. In short, over 120 variables were coded for. Those variables fell into one of the following categories: (a) demographic characteristics, (b) type of article, (c) type of methodology used, (d) type of research design used, (e) independent variables examined, (f) dependent and mediating measures examined, (g) moderating variables examined, (h) measures used, and (1) statistical practices.

Calculating Final Reliabilities

According to Neuendorf (2002), a reliability subsample of between 50 and 200 units is appropriate for estimating levels of interrater agreement. In this case, a simple random reliability subsample of 53 articles was drawn from the sample of 352 articles. Those 53 articles were coded independently by the interrater reliability reviewer so that interrater reliabilities could be estimated.

Because the marginal amounts of each level of variables to be coded were not fixed, Brennan and Prediger's (1981) free-marginal kappa (κm) was used as the statistic of interrater agreement. (By

fixed, we mean that there was not a fixed number of articles that must be assigned to given categories. The marginal distributions were free. See Brennan & Prediger, 1981). Values of kappa lower than .4 were considered to be unacceptable, values between .4 and .6 were considered to be poor, values between and including .6 and .8 were considered to be fair, and values above .8 were considered to be good reliabilities. Confidence intervals around kappa were found through resampling. The resampling code that was used for creating confidence intervals around κm can be found in Appendix D of Randolph (2007a)

Data Analysis

To answer the primary research question, we reported frequencies for each of the multinomial variables or groups of binominal variables. Confidence intervals (95%) for each binary variable or multinomial category were calculated through resampling (see Good, 2001; Simon, 1997), "an alternative inductive approach to significance testing, now becoming more popular in part because of the complexity and difficulty of applying traditional significance tests to complex samples" (Garson, 2006, n.p). The *Resampling Stats* language (1999) was used with the Grosberg's (n.d.) resampling program. Appendix E of Randolph (2007a) presents an example of *Resampling Stats* code that was used to calculate confidence intervals around a proportion.

Results

Interrater Reliability

In short, the interrater reliabilities were good or fair (i.e., greater than .6) (Neuendorf, 2002) for most variables; however, they were lower than .60 on six variables: Kinnunen's (n.d.) categories; type of paper, if not dealing with human participants; literature review present; setting adequately described; procedure adequately described; and results and discussion separate. Five out of seven variables with low reliabilities concern report elements. Specific information about the number of cases (out of 53) that could be used to calculate an interrater reliability statistic, the κm, and its 95% confidence intervals can be found in Randolph (2007a).

Article Demographics

Table 1 shows the numbers of articles that were published in each forum each year. When aggregating the forums into journals or conference proceedings, 289 (76.4%) were published in conference proceedings and 83 (23.6%) were published in journals. (In this case, *Bulletin, CSE*, and *JCSE* were considered to be journals and the other forums were considered to be conference proceedings.) The majority of articles have first authors whom are affiliated with organizations in the U.S. or Canada; see Table 2.

Table 2. Proportions of Regions of First Author's Affiliation

Region	*n*	%	Lower CI 95%	Upper CI 95%
North America	195	55.4	52.3	58.8
Europe	72	20.5	17.6	23.3
Asian-Pacific or Eurasia	50	14.2	12.2	16.2
Middle East	26	7.4	5.2	9.7
Africa	6	1.7	0.1	2.8
South or Central America	3	0.9	0.0	0.0
Total	**352**	**100.0**		

The first author whose articles were most frequently selected in this random sample was Ben-David Kollikant, with four articles. Other first authors whose articles were also frequently selected were A.T. Chamillard, Orit Hazzan, David Ginat, H. Chad Lane, and Richard Rasala, each with three articles in the sample.

The authors of the articles in the selected sample represented 242 separate institutions. Of those 242 institutions, 207 were universities or colleges; 24 were technical universities, institutes of technology, or polytechnics; and 11 were other types of organizations, like research and evaluation institutes or centers. The two institutions with the most articles included in the random sample were the University of Joensuu, with 13 articles or about 4% of the total, and the Technion-Israel Institute of Technology, with six articles or about 2% of total. (Note that 11 of the 13 University of Joensuu articles came from a conference that they help organize.) Other institutions who contributed a large number of articles were Drexel University, Northeastern University, Tel-Aviv University, and the Weismann Institute of Technology, each with five articles in the sample.

The median number of authors on each of the 352 articles was two, with a minimum of one and a maximum of seven. The 2.5^{th} and 97.5^{th} percentiles of the median from 100,000 samples of size 352 were 5 and 5. Of the 349 articles that had page numbers, the median number of pages in the sample was 5, with a minimum of 3 and a maximum of 37. The 2.5^{th} and 97.5^{th} percentiles of the median were 5 and 5.

Report Elements

Table 3 shows the proportion of articles that had or did not have report elements that are considered by the American Psychological Association to be needed in empirical papers that report on investigations with human participants. Note that the interrater reliabilities for the literature review present, purpose/rationale stated, setting adequately described, procedure adequately described, and results and discussion separate variables were low.

Table 3. Proportions of Report Elements

Report element	*n* (of 123)	%	Lower CI 95%	Upper CI 95%
Abstract present	122	99.2	98.4	100.0
Problem is introduced	119	96.7	94.3	99.2
Literature review present	89	72.4	65.9	78.1
Purpose/rationale stated	45	36.6	30.8	42.3
Research questions/hypotheses stated	27	22.0	16.3	27.6
Participants adequately described	56	45.5	39.0	52.0
Setting adequately described	79	64.2	58.5	69.9
Instrument adequately described[a]	66	58.4	52.2	64.6
Procedure adequately described	46	37.4	30.9	43.9
Results and discussion separate	36	29.3	23.6	35.0

Note. Column marginals do not sum to 144 (or 100%) because more than one methodology type per article was possible.
[a] Of 113.

Kinnunen's Content Categories

Table 4 shows how the articles were distributed according to Kinnunen's (n.d.) categories for describing the content of computer science education articles. It shows that the most frequently occurring type of content had to do with a new way to organize a course. Note that the interrater reliability for this variable was poor.

Table 4. Proportions of Articles Falling into Each of Kinnunen's Categories

Content category	*n*	%	Lower CI 95%	Upper CI 95%
New way to organize a course	175	49.7	45.7	54.0
Tool	66	18.8	15.3	22.2
Other	56	15.9	13.1	19.0
Teaching programming languages	31	8.8	6.5	11.4
Parallel computing	10	2.8	1.4	4.3
Curriculum	6	1.7	0.6	2.8
Visualization	6	1.7	0.6	2.8
Simulation	2	0.6	0.0	1.1
Total	**352**	**100.0**		

Valentine's Research Categories

Table 5 shows how the sampled articles were distributed into Valentine's research categories. Experimental and Marco Polo were the most frequently seen types of articles.

Table 5. Proportions of Articles Falling into Each of Valentine's Categories

Valentine's category	n	%	Lower CI 95%	Upper CI 95%
Experimental	144	40.9	36.7	44.9
Marco Polo	118	33.5	29.7	37.5
Tools	44	12.5	9.7	15.3
Philosophy	39	11.1	8.5	13.6
Nifty	7	2.0	0.9	3.1
John Henry	0	0.0		
Total	**352**	**100.0**		

Human Participants

Of the 352 articles in this sample, the majority of articles dealt with human participants. See Table 6.

Table 6. Proportion of Articles Dealing with Human Participants

Human participants	*n*	%	Lower CI 95%	Upper CI 95%
Yes	233	66.2	62.2	70.1
No	119	33.8	29.8	37.8
Total	**352**	**100.0**		

Grade Level of Participants

Table 7 shows the grade level of participants of the 123 articles that dealt with human participants, that were not qualitative only, and that presented more than anecdotal evidence (hereafter these 123 articles are called *the behavioral, quantitative, and empirical articles*). Bachelor's degree students were overwhelmingly the type of participants most often investigated in the articles in this sample.

Table 7. Proportions of Grade Level of Participants

Grade level of participant	*n*	%	Lower CI 95%	Upper CI 95%
Preschool	2	2.3	0.0	5.7
K-12	5	5.7	2.3	10.2
Bachelor's level	64	72.7	64.8	80.7
Master's level	1	1.1	0 .0	3.4
Doctoral level	0	0.0		
Mixed level/other	16	18.2	11.4	25.0
Total	**88**	**100.0**		

As Table 8 shows, of the 64 Bachelor's degree participants, most were taking first-year computer science courses at the time the study was conducted. Studies in which the participants were not students (e.g., teachers) or the participants were of mixed grade levels were included in the mixed

level/other category. (Note that the interrater reliability for the grade level of participants' variable, but not the undergraduate year variable, was below a kappa of .4).

Table 8. Proportion of Undergraduate Level of Computing Curriculum

Year of undergraduate level computing curriculum	*n*	%	Lower CI 95%	Upper CI 95%
First year	39	70.9	61.8	80.0
Second year	3	5.5	1.8	90.9
Third year	8	14.5	7.3	2.2
Fourth year	5	9.1	3.6	14.6
Total	**64**	**100.0**		

Anecdotal Evidence Only

Of the 233 articles that dealt with human participants, 38.2% presented only anecdotal evidence. See Table 9.

Table 9. Proportion of Human Participants Articles that Provided Anecdotal Evidence Only

Anecdotal	*n*	%	Lower CI 95%	Upper CI 95%
Yes	89	38.2	33.1	43.3
No	144	61.8	56.7	66.5
Total	**233**	**100.0**		

Types of Articles That Did Not Deal with Human Participants

Of the 119 articles that did not deal with human participants, the majority were purely descriptions of interventions. See Table 10, which shows the proportions of those articles that were program descriptions; theory, methodology, or philosophical papers; literature reviews; or technical papers. (Note that the interrater reliability estimate of kappa for this variable was below .6.)

Table 10. Proportions of Types of Articles Not Dealing With Human Participants

Type of article	*n*	%	Lower CI 95%	Upper CI 95%
Program description	72	60.5	53.8	67.2
Theory, methodology, or philosophical paper	36	30.3	24.4	37.0
Literature review	10	8.4	5.0	11.8
Technical	1	0.8	0.0	1.7
Total	**119**	**100.0**		

Types of Research Methods Used

Table 11 shows that the experimental/quasi-experimental methodology type was the most frequently used type of methodology in the articles that dealt with human participants and that pre-

sented more than anecdotal evidence. Table 12 shows the proportions of quantitative articles, qualitative articles, and mixed-methods articles.

Table 11. Proportion of Methodology Types Used

Methodology type	*n* (of 144)	%	Lower CI 95%	Upper CI 95%
Experimental /quasi-experimental	93	64.6	58.3	70.8
Qualitative	38	26.4	20.8	31.3
Causal comparative	26	18.1	13.2	22.9
Correlational	15	10.4	7.0	14.6
Survey research	11	7.6	4.2	11.1

Note. Column marginals do not sum to 144 (or 100%) because more than one methodology type per article was possible.

Table 12. Proportion of Types of Methods

Type of Method	*n*	%	Lower CI 95%	Upper CI 95%
Quantitative	107	74.3	68.1	80.2
Qualitative	22	15.3	10.4	20.8
Mixed	15	10.4	6.3	14.6
Total	**144**	**100**		

In terms of the 144 studies that dealt with human participants and that presented more than anecdotal evidence, convenience sampling of participants was used in 124 (86.1%) of the cases, purposive (nonrandom) sampling was used in 14 (9.7%) of the cases. Random sampling was used in 6 (4.2%) of the cases.

Research Designs

Table 13 shows that the most frequently used research design was the one-group posttest-only design (i.e., the ex post facto design). Of the 51 articles that used the one-group posttest-only design, 46 articles used it exclusively (i.e., they did not use a one-group posttest-only design *and* a research design that incorporated a pretest or a control of contrast group).

Table 13. Proportions of Types of Experimental/Quasi-Experimental Designs Used

Type of experimental design	*n* (of 93)	%	Lower CI 95%	Upper CI 95%
Posttest only	51	54.8	47.3	62.4
Posttest with controls	22	23.7	17.2	30.1
Pretest/posttest without controls	12	12.9	8.6	18.3
Repeated measures	7	7.5	4.3	11.8
Pretest/posttest with controls	6	6.5	2.2	10.8
Single-subject	3	3.2	1.1	5.3

Note. Column marginals do not sum to 93 (or 100%) because more than one methodology type per article was possible.

In the sampled articles, quasi-experimental studies were much more frequently conducted than truly experimental studies. Of the 93 studies that used an experimental or quasi-experimental methodology, participants self-selected into conditions in 81 (87.1%) of the studies, participants were randomly assigned to conditions in 7 (7.5%) of the studies, and participants were assigned to conditions purposively, but not randomly, by the researcher(s) in 5 (5.4%) of the studies.

Independent Variables

Table 14 shows the proportions of types of independent variables that were investigated in the 93 articles that used an experimental/quasi-experimental methodology. Nearly 99% of all independent variables were related to student instruction.

Table 14. Proportion of Types of Independent Variables Used

Type of independent variable used	*n* (of 93)	%	Lower CI 95%	Upper CI 95%
Student instruction	92	98.9	96.8	1.0
Teacher instruction	4	4.3	2.2	6.5
Mentoring	2	2.2	0.0	5.3
Speakers at school	2	2.2	0.0	5.3
Field trips	1	1.1	0.0	2.2
Computer science fair/contest	0	0.0		

Note. Column marginals do not sum to 93 (or 100%) because more than one type of independent variable could have been used in each article (e.g., when there were multiple experiments).

Dependent Variables

Table 15 shows the proportions of the different types of dependent variables that were measured in the 123 behavioral, quantitative, and empirical articles. Table 15 shows that attitudes and achievement in computer science were the dependent variables that were most frequently measured. The variables *project implementation* and *costs and benefits*, although included as categories on the coding sheet, are not included in Table 15 because there were no studies that used them as dependent measures.

Table 15. Proportions of Types of Dependent Variables Measured

Type of dependent variable measured	*n* (of 123)	%	Lower CI 95%	Upper CI 95%
Attitudes (student or teacher)	74	60.2	53.7	66.7
Achievement in computer science	69	56.1	49.6	62.6
Attendance	26	21.1	15.5	28.3
Other	14	11.5	7.4	15.6
Computer use	5	4.1	1.6	6.5
Students' intention for future	3	2.4	0.1	4.9
Teaching practices	2	1.6	0.0	3.3
Achievement in core (non-cs) courses	1	0.8	0.0	2.4
Socialization	1	0.8	0.0	2.4

Note. Column marginals do not sum to 123 (or 100%) because more than one type of dependent variables could have been measured.

Mediating or Moderating Variables

Of the 123 behavioral, quantitative, and empirical articles, moderating or mediating variables were examined in 29 (23.6%). Table 16 shows the types and proportions of moderating or mediating variables that were examined in the sample of articles. There were many articles that examined moderating or mediating variables that fit into the *other* category (i.e., they were not originally on the coding sheet); those other variables were tabulated and have been incorporated into Table 17. Although included on the coding sheet, the variables—*disability* and *socioeconomic status*—were not included in Table 16 because no study examined them as mediating or moderating variables.

Table 16. Proportions of Mediating or Moderating Variables Investigated

Mediating or moderating variable investigated	n (of 29)	%	Lower CI 95%	Upper CI 95%
Gender	6	20.7	13.8	27.6
Grade level[a]	4	13.8	6.9	20.7
Learning styles[a]	4	13.8	6.9	20.7
Aptitude (in computer science)[a]	2	6.8	3.5	10.3
Major/minor subject[a]	2	6.8	3.5	10.3
Race/ethnic origin	2	6.8	3.5	10.3
Age[a]	1	3.4	0.0	6.9
Amount of scaffolding provided[a]	1	3.4	0.0	6.9
Frequency of cheating[a]	1	3.4	0.0	6.9
Pretest effects[a]	1	3.4	0.0	6.9
Programming language[a]	1	3.4	0.0	6.9
Type of curriculum[a]	1	3.4	0.0	6.9
Type of institution[a]	1	3.4	0.0	6.9
Type of computing laboratory[a]	1	3.4	0.0	6.9
Type of grading (human or computer[a])	1	3.4	0.0	6.9
Self-efficacy[a]	1	3.4	0.0	6.9

Note. Column marginals do not sum to 29 (or 100%) because more than one methodology type per article was possible.

[a] These items were not a part of the original coding categories.

Types of Measures Used

Table 17 shows the proportions of types of measures that were used in the 123 behavioral, quantitative, and empirical articles. Note that questionnaires were clearly the most frequently used type of measure. Measurement validity or reliability data were provided for questionnaires in 1 of 65 (1.5 %) of articles, for teacher- or researcher-made tests in 5 of 27 (18.5 %) of articles, for direct observation (e.g., interobserver reliability) in 1 of 4 (25%) of articles, and for standardized tests in 6 of 11 (54.5%) of articles.

Table 17. Proportions of Types of Measures Used

Type of measure used	*n* (of 123)	%	Lower CI 95%	Upper CI 95%
Questionnaires	65	52.8	46.3	59.4
Grades	36	29.3	23.6	35.0
Teacher- or researcher-made tests	27	22.0	16.3	27.6
Student work	22	17.9	13.0	23.6
Existing records	20	16.3	11.4	21.1
Log files	15	12.2	8.1	9.2
Standardized tests	11	8.9	4.9	13.0
Interviews	8	6.5	3.3	9.8
Direct observation	4	3.3	0.8	5.7
Learning diaries	4	3.3	0.8	5.7
Focus groups	3	2.4	0.8	4.9

Note. Column marginals do not sum to 123 because more than one measure per article was possible.

Type of Inferential Analyses Used

Of the 123 behavioral, quantitative, and empirical articles, inferential statistics were used in 44 (35.8%) of them. The other 79 articles reported quantitative results, but did not use inferential analyses. Table 18 shows the types of inferential statistics used, their proportions, and the proportion of articles that provided statistically adequate information along with the inferential statistics that were reported.

Table 18. Proportions of Types of Inferential Analyses Used

Type of inferential analysis used	*n*	%	Lower CI 95%	Upper CI 95%
Parametric analysis (of 44)	25	56.8	47.7	65.9
Measure of centrality and dispersion reported (of 25)	15	60.0	48.0	72.0
Correlational analysis (of 44)	13	29.5	23.3	37.2
Sample size reported (of 13)	10	76.9	53.9	92.3
Correlation or covariance matrix reported (of 13)	5	38.5	15.4	61.5
Nonparametric analysis (of 44)	11	25.0	13.2	31.8
Raw data summarized (of 11)	8	72.7	45.6	90.9
Small sample analysis (of 44)	2	4.5	0.0	9.1
Entire data set reported (of 2)	0	0.0		
Multivariate analysis (of 44)	1	2.3	0.0	2.3
Cell means reported (of 1)	0	0.0		
Cell sample size reported (of 1)	0	0.0		
Pooled within variance or covariance matrix reported (of 1)	0	0.0		

Note. Column marginals do not sum because more than one methodology type per article was possible.

Type of Effect Size Reported

Of the 123 behavioral, quantitative, and empirical articles, 120 (97.6%) reported some type of effect size. In the three articles that reported quantitative statistics but not an effect size, those articles presented only probability values or only reported if the result was "statistically significant" or not. Table 19 presents the types of effect sizes that were reported and their proportions. Odds, odds ratio, or relative risk were not reported in any of the articles in this sample. Of the articles that reported a raw difference effect size, 74 of those reported the raw difference as a difference between means (the rest were reported as raw numbers, proportions, means, or medians). Of the 74 articles that reported means, 29 (62.5%) did not report a measure of dispersion along with the mean. Note that a liberal definition of a raw difference—also referred to as relative risk or a gain score—was used here. The authors did not actually have to subtract pretest and posttest raw scores (or pretest and posttest proportions) from one another to be considered a raw difference effect size. They simply had to report two raw scores in such a way that a reader could subtract one from another to get a raw difference.

Table 19. Proportions of Types of Effect Sizes Reported

Type of effect size reported	*n* (of 120)	%	Lower CI 95%	Upper CI 95%
Raw difference	117	97.5	95.0	100.0
Correlational effect size	8	6.7	3.3	6.7
Standardized mean difference	6	5.0	1.7	8.3

Note. Column marginals do not sum to 120 (or 100%) because more than one methodology type per article was possible. About one third of articles did not report research on human participants.

Discussion

Study Limitations

One study limitation was that the interrater reliabilities were low on a small proportion of the variables. We tried to circumvent this study limitation by not making strong conclusions about variables with poor reliabilities or by qualifying claims that were supported by variables with poor reliabilities.

As was mentioned in the Methods section, we recognize that we approached this review from the viewpoint of primarily quantitatively oriented behavioral science researchers. We investigated most deeply the quantitative experimental articles and did not deeply analyze articles that exclusively used qualitative modes of inquiry. Because of the significant variety and variability of qualitative methods, we were not confident that we could develop (or implement) a reliable system of classifying, analyzing, and evaluating those articles. Therefore, another study limitation was that we concentrated on experimental articles at the expense of qualitative articles.

Revisiting Research Questions

Our primary research question, which we addressed in terms of nine sub-questions, was, "What are the methodological properties of research reported in articles in major computer science education research forums from the years 2000-2005?" A short answer to each of those research subquestions is dealt with below.

What was the proportion of articles that did not report research on human participants?

We found that about one-third of the articles did not report research on human participants. Those articles were literature reviews, theoretical or methodological articles, program descriptions, etc. The proportion in the current review (33.8%) was about 30% lower than in the Randolph, Bednarik, & Myller (2005) review of the articles in the proceedings of the Koli Calling conferences.

Of the articles that did not report research on human participants, what types of articles were being published and in what proportions?

Of the 34% of papers that did not report research on human participants, most (60%) of the papers were purely descriptions of interventions without any analysis of the effects of the intervention on computer science students. This proportion of articles is slightly higher, but near, the proportion of program descriptions in other computing-related methodological reviews in which the proportion of program descriptions was measured. Assuming that Valentine's (2004) categories Marco Polo and Tools coincide with our program description category, then Valentine's findings are similar to our own; he found that 49% of computer science education research articles are what he called Marco Polo or Tools articles. Similarly, Tichy, Lukowicz, Prechelt, and Heinz (1995) found that 43% of the computer science articles in their study were design and modeling articles, which would be called program descriptions in our categorization system.

Of the articles that did report research on human participants, what proportion provided only anecdotal evidence for their claims?

The issue of the proliferation of anecdotal evidence in computing research, especially in software engineering, was being addressed over ten years ago. Holloway (1995) wrote, "rarely, if ever, are [empirical claims about software engineering] augmented with anything remotely resembling either logical or empirical evidence. . . resting an entire discipline on such a shaky epistemological foundation is absurd, but ubiquitous nonetheless" (p. 21).

As Table 9 showed, the proliferation of anecdotal evidence is also an issue for the current computer science education research. Note that by the term anecdotal evidence in this review we have meant the informal observation of a phenomenon by a researcher. We do not necessarily mean that humans cannot make valid and reliable observations themselves, as happens in ethnographic research or research in which humans operationalize and empirically observe behavior. Also, we concur that anecdotal experience has a role in the research process—it has a role in hypothesis generation. But, as Holloway (1995) pointed out, there are major problems to using informal anecdotal experience as the sole means of hypothesis confirmation. Valentine (2004) in his methodological review came to the same conclusion about the proliferation of anecdotal evidence in the field computer science education research. This sentiment about the importance of collecting empirical data is also echoed in several papers on computer science education research such as Clancy, Stasko, Guzdzial, Fincher, and Dale (2001) and Holmboe, McIver, and George (2001).

Of the articles that did report research on human participants, what types of methods were used and in what proportions?

Experimental investigations were reported in nearly 65% of computer science education articles reviewed here (see Table 11). However, as we explain later, the experimental designs that were predominantly used were prone to almost all threats to internal validity. After experimental methods, qualitative methods were the next most frequently used methods.

Experimental/quasi-experimental and qualitative methods are both methods that allow researchers to make causal inferences, and thereby confirm their causal hypotheses (Mohr, 1999). Experimental/quasi-experimental research is predicated on a comparison between a counterfactual and

factual condition, via, what Mohr called, factual causal reasoning. Qualitative research is predicated on what Mohr called physical causal reasoning, or what Scriven (1976) called the Modus Operandi Method of demonstrating causality. At any rate, the fact that most of the research being done in computer science education is done with types of methods that could possibly arrive at causal conclusions (given that the research is conducted properly) is a positive sign for computer science education research.

Of the articles that did report research on human participants, what measures were used, in what proportions, and was psychometric information reported?

Questionnaires were clearly the most frequently used type of measure. In fact, as Table 17 shows, over half of the measures were questionnaires. Grades and teacher- or researcher-made tests were the second and third most commonly used measures, respectively.

One alarming finding was that only 1 out of 65 articles in which questionnaires were used gave any information about the reliability or validity of the instrument. According to Wilkinson and the Task Force on Statistical Inference, "if a questionnaire is used to collect data, [a researcher should] summarize the psychometric properties of its scores with specific regard to the way the instrument is used in a population. Psychometric properties include measures of validity, reliability and internal validity" (1999, n.p). Obviously, the lack of psychometric information about instruments is a clear weakness in the body of the computer science education research.

Of the articles that did report research on human participants, what were the types of independent, dependent, mediating, and moderating factors that were examined and in what proportions?

Mark Guzdzial, one of the members of the working group on Challenges to Computer Science Education Research, admits that, "We know that student opinions are unreliable measures of learning or teaching quality" (Almstrum et al., 2005, p. 191). Yet, this review shows that attitudes are the most frequently measured variable. In fact, 44% of articles used attitudes as the sole independent article. While attitudes may be of interest to computer science education researchers, as Guzdzial suggests, they are unreliable indicators of learning or teaching quality.

Of the articles that used experimental/quasi-experimental methods, what types of designs were used and in what proportions?

It is clear that the one-group posttest-only and posttest-only with control designs were the most frequently used types of experimental research designs. It is important to note that the one-group posttest-only design was used more than twice as often as the next most frequently used design, the posttest-only design with controls. Other designs, with pretests and/or control groups, obviously would have been better design choices if the goal had been causal inference. The one-group post-test design is subject to almost all threats to internal validity (Shadish, Cook, & Campbell, 2002).

In terms of selection and assignment procedures, we found convenience samples were used in 86% of articles, and students self-selected into treatment and control conditions in 87% of the articles. While some, such as Kish (1987) and Lavori, Louis, Bailar, and Polansky (1986), are staunch advocates of the formal model of sampling (i.e., random sampling followed by random assignment), there are others that question that model's utility. Others, such as Shadish and colleagues (2002) and Wilkinson and the Task Force on Statistical Inference (1999), claim that formal sampling methods have limited utility. The debate is ongoing.

The conclusion for computer science education researchers is that while random sampling is desirable when it can be done, doing purposive sampling or at least assessing the representativeness of a sample by examining surface similarities, ruling out irrelevancies, making discriminations,

interpolating and extrapolating, and examining causal explanations can be a reasonable alternative.

In terms of random assignment of participants to treatment conditions, the same types of lessons apply. While random assignment is desirable, when it is not feasible there are other ways to make strong causal conclusions. When it is not possible to randomly assign participants to experimental conditions, steps need to be made, through design or analysis, to measure and then minimize the effects of confounding variables: "variables that affect the observed relations between a causal variable and an outcome" (Wilkinson & Task Force on Statistical Inference, 1999, n.p.). For example, one might measure the previous computing experience of participants and then use that information to statistically control for previous computing experience.

Of the articles that reported quantitative results, what kinds of statistical practices were used and in what proportions?

The American Psychological Association (2001, p. 23) suggests that certain information be provided when certain statistical analyses are used. For example, when parametric tests of location are used "a set of sufficient statistics consists of cell means, cell sample sizes, and some measures of variability. . . . Alternately, a set of sufficient statistics consists of cell means, along with the mean square error and degrees of freedom associated with the effect being tested." Second, the American Psychological Association (2001) and the American Psychological Association's Task Force on Statistical Inference Testing (Wilkinson & Task Force on Statistical Inference, 1999) argue that it is best practice to report an effect size in addition to *p*-values.

The results of this review showed that inferential analyses are conducted in 36% of cases when quantitative results are reported. When computer science educators do conduct inferential analyses, only a moderate proportion report informationally adequate statistics. Areas of concern include reporting a measure of centrality *and* dispersion for parametric analyses, reporting sample sizes and correlation or covariance matrices for correlational analyses, and summarizing raw data when nonparametric analyses are used.

Of the articles that did report research on human participants, what were the characteristics of the articles' structures?

There were several interesting findings about the elements of the papers reviewed here. For example, about 25% of empirical articles were missing a literature review, 22% had no stated research questions, and less than 50% of articles adequately described instruments or procedures. However, we are not confident about making a strong claim about the presence or absence of literature reviews in the articles in the current review because of the low levels of interrater agreement on the variables relating to reporting elements. However, we think that the fact that two raters could not reliably agree on the presence or absence of key report elements; such as the literature review, research questions, report elements, description of participants, description of procedure; at least points out that these elements need to be explained more clearly. For example, if two raters cannot agree on whether or not there is a literature review in an academic paper, we are inclined to believe that the literature review is flawed in some way.

Assuming that the literature reviews in computer science education research articles are indeed lacking, then it is no surprise that the ACM SIGCSE Working Group on Challenges to Computer Science Education concluded that there is a lack of accumulated evidence and a tendency for computer science educators to "reinvent the wheel" (Almstrum et al., 2005, p. 191). Besides allowing evidence to accumulate and not reinventing the wheel, conducting thorough literature reviews takes some of the burden off researchers who are attempting to gather evidence for a claim since "good prior evidence often reduces the quality needed for later evidence" (Mark, Henry, & Julnes, 2000, p. 87).

Also, one conclusion that can be drawn from the fact that the literature review and other report elements variables had such low reliabilities is that the traditions of reporting differ significantly between what is suggested by the American Psychological suggestion and how most computer science education reports are structured. While not having agreed upon structures enables alternative styles of reporting to flourish and gives authors plenty of leeway to present their results, it makes it difficult for the reader to quickly extract needed information from the articles. Additionally, we hypothesize that the lack of agreed upon structures for computer science education articles leads to the omission of critical information needed in reports of research with human participants, such as a description of procedures and participants, especially by beginning researchers. Note that the report element variables, such as the lack of a literature review or the lack of information about participants or procedures, only pertained to articles that reported on investigations with human participants and not to other types of articles, such as program descriptions or theoretical papers, in which the report structures would obviously differ from a report of an investigation with human participants.

Recommendations

In this section we report on what we consider to be the most important evidence-based recommendations for improving the current state of computer science education. Because we expect that the improvements will be most likely effected by editors and reviewers raising the bar in terms of the methodological quality of papers that get accepted for publication, we direct these recommendations primarily to the editors and reviewers of computer science education research forums. Also, these recommendations are relevant to funders of computer science research; to consumers of computer science education research, such as educational administrators; and, of course, to computer science education researchers themselves.

Anecdotal Experience

While a field probably cannot be built entirely on anecdotal experience (although some might not agree), that does not mean that anecdotal experience does not have an important role in scientific inquiry—it has an important role in the generation of hypotheses. Sometimes it is through anecdotal experience that researchers come to formulate important hypotheses. However, because of its informality, anecdotal experience is certainly a dubious type of evidence for hypothesis confirmation. Not accepting anecdotal evidence as a means of hypothesis confirmation is not to say that a human cannot make valid and reliable observations. However, there is a significant difference between a researcher reporting that "we noticed that students learned a lot from our program" and a researcher who reports on the results of a well-planned qualitative inquiry or on the results of carefully controlled direct observations of student behavior, for example. Our recommendation is for reviewers to *accept anecdotal experience as a means of hypothesis generation, but not as a sole means of hypothesis confirmation*.

Self-Reports of Learning

Of course, stakeholders' reports about how much they have learned are important; however, it probably is not the only dependent of variable of interest in an educational intervention. As a measure of learning, as Guzdzial (in Almstrum et al., 2005) has pointed out, students' opinions are poor indicators of how much learning has actually occurred. We recommend that reviewers *be wary of investigations that only measure students' self-reports of learning*.

Reliability and Validity of Measures

Wilkinson and the Task Force on Statistical Inference (1999) provided valuable advice to editors concerning this issue, especially in "a new and rapidly growing research area" (like computer science education). They advised,

> Editors and reviewers should pay special attention to the psychometric properties of the instrument used, and they might want to encourage revisions (even if not by the scale's author) to prevent the accumulation of results based on relatively invalid or unreliable measures. (n.p.)

Our recommendation in this area is to insist that authors provide some kind of information about the reliability and validity of measures that they use.

The One-Group Posttest-Only Design

In the one-group posttest-only design, almost any influence could have caused the result. For example, in a one-group posttest-only design, if the independent variable was an automated tool to teach programming concepts and the dependent variable was the mastery of programming concepts, it is entirely possible that, for example, students already knew the concepts before using the tools, or that something other than the tool (e.g., the instructor) caused the mastery of the concepts. Experimental research designs that compare a factual to a counterfactual condition are much better at establishing causality than research designs that do not. Our recommendation is to *realize that the one-group posttest-only research design is susceptible to almost all threats to internal validity*.

Informationally Adequate Statistics

When inferential statistics are used, be sure that the author includes enough information for the reader to understand the analysis used and to examine alternative hypotheses for the results that were found. The American Psychological Association (2001) provides a good description of what types of information is expected to be reported for certain types of statistical analyses. We recommend that reviewers *make sure that authors report informationally adequate statistics*.

Detail about Participants and Procedures

When authors report research on human participants be sure that they include adequate information about the participants, apparatus, and procedure. See American Psychological Association (2001) for guidelines on what is considered to be sufficient detail in describing participants and procedures. In short, enough information should be provided about participants so that readers can determine generalization parameters and enough information should be provided about the procedure that it could be independently replicated. Our final recommendation is for reviewers *to insist that authors provide sufficient detail about participants and procedures*.

Conclusion

Summary

In this methodological review, we used a content analysis approach to conduct a methodological review of the articles published in mainstream computer science education forums from 2000 to 2005. Of the population of articles published during that time a random sample of 352 articles was drawn; each article was reviewed in terms of its general characteristics; the type of methods used; the research design used; the independent, dependent, and mediating or moderating vari-

ables used; the measures used; and statistical practices used. The major findings from the review are listed below:

- About one third of articles did not report research on human participants.
- Most of the articles that did not deal with human participants were program descriptions.
- Nearly 40% of articles that dealt with human participants only provided anecdotal evidence for their claims.
- Of the articles that provided more than anecdotal evidence, most articles used experimental/quasi-experimental or qualitative methods.
- Of the articles that used an experimental research design, the majority used a one-group posttest-only design exclusively.
- Student instruction, attitudes, and gender were the most frequent independent, dependent, and mediating/moderating variables, respectively.
- Questionnaires were clearly the most frequently used type of measurement instrument. Almost all of the measurement instruments that should have psychometric information provided about them did not have psychometric information provided.
- When inferential statistics were used, the amount of statistical information used was inadequate in many cases.

Based on these findings, we made the following recommendations to editors, reviewers, authors, funders, and consumers of computer science education research:

- Accept anecdotal experience as a means of hypothesis generation, but not as the sole means of hypothesis confirmation.
- Be wary of investigations that measure only students' attitudes and self-reports of learning as a result of an intervention.
- Insist that authors provide some kind of information about the reliability and validity of measures that they use.
- Realize that the one-group posttest-only research design is susceptible to almost all threats to internal validity.
- Encourage authors to report informationally adequate statistics.
- Insist that authors provide sufficient detail about participants and procedures.

Computer Science Education Research at the Crossroads

Based on the results of this review, we can conclude what computer science educators have excelled at is generating a large number of informed research hypotheses based on anecdotal experience or on poorly designed investigations. However, they have not systematically tested these hypotheses. This leaves computer science education at a crossroads. To the crossroads computer science education researchers bring a proliferation of well-informed hypotheses. What will happen to these hypotheses remains to be seen.

One option is that these informed hypotheses will overtime, through repeated exposure, "on the basis of 'success stories' and slick sales pitches" (Holloway, 1995, p. 20) come to be widely accepted as truths although having never been empirically verified. That is, they will become folk conclusions. (We use the term *folk conclusions* instead of *folk theorems* (See Harel, 1980) or *folk*

myths (See Denning, 1980) since the validity of the conclusion has not yet been empirically determined.)

Because scientific knowledge usually develops cumulatively, if informed hypotheses are allowed to developed into folk conclusions, then layers of folk conclusions (both true and untrue) will become inexorably embedded in the cumulative knowledge of what is known about computer science education. Computer science education will become a field of research whose foundational knowledge is based on conclusions that are believed to be true, but which have never been empirically verified. Indeed, as Holloway suggests "resting an entire discipline on such a shaky epistemological foundation is absurd . . ." (1995, p. 21). In the same vein, basing the future of an entire discipline on such a shaky epistemological foundation is also absurd.

We are not arguing, however, that hypothesis generation or any other type of research activity in computer science education should be abandoned altogether. There needs to be a requisite variety of methods to draw from so that a rich variety of research acts can be carried out. Also, hypothesis generation is inexorably tied with innovation.

What we are arguing is that the proportions of research methods being used needs to be congruent with the current challenges and problems in computer science education. If the ACM SIGCSE's Working Group on Challenges to Computer Science Education is correct that the current challenges involve a lack or rigor and accumulated evidence (Almstrum et al., 2005), then it makes sense to shift the balance from one that emphasizes anecdotal evidence and hypothesis generation to one that emphasizes rigorous methods and hypothesis confirmation. Coming back to the discussion of the crossroads, the sustainable path for computer science education involves building on the hypotheses of the past and striking a balance between innovation and experimentation in the future.

Acknowledgments

This work is based on/excerpted from the first author's doctoral dissertation (Randolph, 2007a) accepted for publication by UMI Proquest and VDM Verlag. This research was supported in part by a generous Special Projects Grant from ACM SIGCSE. Thanks to Jim Doward and Stephen Clyde for their comments on the work on which this article was based.

References

Almstrum, V. L., Hazzon, O., Guzdzial, M., & Petre, M. (2005). Challenges to computer science education research. In *Proceedings of the 36th SIGCSE Technical Symposium on Computer Science Education SIGCSE '05* (pp.191-192). New York: ACM Press.

American Psychological Association. (2001). P*ublication manual of the American Psychological Association* (5th ed). Washington, DC: American Psychological Association.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41,* 687-699.

Clancy, M., Stasko, J., Guzdzial, M., Fincher, S., & Dale, N. (2001). Model and areas for CS education research. *Computer Science Education, 11*(4), 323-341.

Denning, P. J. (1980). On folk theorems, and folk myths. *Communications of the ACM, 23*(9), 493-494.

Denning, P. J., Comer, D. E., Gries, D., Mulder, M. C., Tucker, A. Turner, A. J., et al. (1989). Computing as a discipline. *Communications of the ACM, 32*(1), 9-23.

Fincher, S., & Petre, M. (2004). *Computer science education research.* London: Taylor & Francis.

Garson, D. V. (2006). *Sampling*. Retrieved March 28, 2006, from North Carolina State University, College of Humanites & Social Science Web site: http://www2.chass.ncsu.edu/garson/PA765/sampling.htm

Good, P. I. (2001). *Resampling methods. A practical guide to data analysis* (2nd ed.). Boston: Birkhäuser.

Grosberg, J. (n.d.). *Statistics 101* [Computer Software]. Retrieved July 11, 2006, from http://www.statistics101.net/index.htm

Harel, D. (1980). On folk theorems. *Communications of the ACM, 23*(7), 379-494.

Holloway, C. M. (1995). Software engineering and epistemology. *Software Engineering Notes, 20*(2), 20-21.

Holmboe, C., McIver, L., & George, C. (2001). Research agenda for computer science. In *Proceedings of the 13th Annual Workshop of the Psychology of Programming Interest Group* (pp. 207-223).

Kinnunen, P. (n.d.) *Guidelines of computer science education research*. Retrieved November 29, 2005, from http://www.cs.hut.fi/Research/COMPSER/ROLEP/seminaari-k05/S_05-nettiin/Guidelines_of_CSE-teksti-paivi.pdf

Kish, I. (1987). *Statistical design for research.* New York, NY: Wiley.

Lavori, P. W., Louis, T. A., Bailar, J. C., & Polansky, H. (1986). Design of experiments: Parallel comparisons of treatments. In J. C. Bailar & F. Mosteller (Eds.), *Medical uses of statistics* (pp. 61-82). Waltham, MA: New England Journal of Medicine.

Mark, M. M., Henry, G. T., & Julnes, G. (2000). *Evaluation: An integrated framework for understanding, guiding, and improving policies and programs*. San Francisco, CA: Jossey-Bass.

Mohr, L. B. (1999). The qualitative method of impact analysis. *American Journal of Evaluation, 20*(1), 69-84.

Neuendorf, K. A. (2002). *The content analysis handbook.* Thousand Oaks, CA: Sage.

Randolph, J. J. (2007a). *Computer science education research at the crossroads: A methodological review of the computer science education research: 2000-2005.* Unpublished dissertation, Utah State University, Logan, Utah. Retrieved April 24, 2008, from http://www.archive.org/details/randolph_dissertation

Randolph, J. J. (2007b). What's the difference, still: A follow-up review of the quantitative research methodology in distance learning. *Informatics in Education, 6*(1), 179-188.

Randolph, J. J. (in press). A methodological review of the program evaluations in K-12 computer science education. *Informatics in Education.*

Randolph, J. J., Bednarik, R., & Myller, N. (2005). A methodological review of the articles published in the proceedings of Koli Calling 2001-2004. In *Proceedings of the 5th Annual Finnish / Baltic Sea Conference on Computer Science Education* (pp. 103-109). Finland: Helsinki University of Technology Press.

Randolph, J. J., Bednarik, R., Silander, P., Lopez-Gonzalez, J., Myller, N., & Sutinen, E. (2005). A critical review of research methodologies reported in the full-papers of ICALT 2004. In *Proceedings of the Fifth International Conference on Advanced Learning Technologies* (pp.10-14). Los Alamitos, CA: IEEE Press.

Randolph, J. J., & Hartikainen, E. (2005). A review of resources for K-12 computer-science-education program evaluation. In *Yhtenäistyvät vai erilaistuvat oppimisen ja koulutuksen polut: Kasvatustieteen päivien 2004 verkkojulkaisu* (Electronic Proceedings of the Finnish Education Research Days Conference 2004) (pp. 183-193). Finland: University of Joensuu Press.

Randolph, J.J., Hartikainen, E., & Kähkönen, E. (2004). Lessons learned from developing a procedure for the critical review of educational technology research. Paper presented at *Kasvatustieteen Päivät 2004* (Finnish Education Research Days Conference 2004), Joensuu, Finland, November, 2004.

Resampling Stats (Version 5.0.2) [Computer software and manual]. (1999). Arlington, VA: Resampling Stats.

Sample Planning Wizard [Computer software]. (2005). Stat Trek.

Scriven, M. (1976). Maximizing the power of causal investigations: The modus operandi method. In G. V. Glass (Ed.), *Evaluation studies review annual, Vol. 1* (pp. 101-118). Beverly Hills, CA: Sage.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Simon, J. L. (1997). *Resampling: The new statistics.* Arlington, VA: Resampling Stats.

Tichy, W. F., Lukowicz, P., Prechelt, L., & Heinz, E. A. (1995). Experimental evaluation in computer science. A quantitative study. *Journal of Systems and Software, 28,* 9-18.

Valentine, D. W. (2004). CS educational research: A meta-analysis of SIGCSE technical symposium proceedings. In *Proceedings of the 35th Technical Symposium on Computer Science Education* (pp 255-259). New York: ACM Press.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations (Electronic version). *American Psychologist, 54,* 594-604.

Biographies

Justus Randolph has a PhD in education research and program evaluation, an MEd in international education, and a certification in educational administration. Currently he works as a researcher for the InnoPlay project at the Centre for Media Pedagogy at the University of Lapland's Faculty of Education. His research and evaluation experiences have concerned programs that involve newborn hearing assessment, school improvement, higher education evaluation, technology-enriched playgrounds, educational technology research methods, and computing education. He is the author of the book *Multidisciplinary Methods in Educational Technology Research and Development.*

George Julnes is an Associate Professor of Psychology at Utah State University, with primary interests in community psychology and the use of program and policy evaluation to improve society. Prof. Julnes received his Ph.D. in psychology (clinical/community) at the University of Hawaii-Manoa in 1984. In support of his further development in program and policy evaluation, he earned subsequent graduate degrees in public policy (MPP) and business (MBA) at the University of Michigan in 1989. He is a co-chair of the Quantitative Methods Topical Interest Group of the American Evaluation Association and serves as a grant and journal reviewer in the general area of evaluation methods and theory.

Erkki Sutinen is a professor of Computer Science and the head of the Department of Computer Science and Statistics at the University of Joensuu, Finland. He got his PhD from the University of Helsinki in the area of string algorithms. Before joining the faculty at Joensuu, he worked at Purdue University, USA, and University of Linköping, Sweden. He has also been a visiting faculty member at Massey University, NZ, and University of Pretoria, South Africa. He has been appointed an adjunct professor at Tumaini University, Tanzania. Professor Sutinen is interested in how new technologies can transform traditional ways and approaches of learning. His research group (www.cs.joensuu.fi/edtech) develops various visualization, story telling, pervasive and robotic tools together with actual users. Contextualizing ICT education for the needs of developing countries and supporting diverse learners from special education have both shown to be living laboratories, leading to R&D initiatives that are relevant also to regular education in conventional settings.

Steve Lehman, an educational psychologist, was appointed as an Associate Professor of Psychology at Utah State University in 2001. His primary research interested includes text processing, reading comprehension, and web-based instruction.