Spectrum of Engineering Sciences
Vol-1, Issue-1,
http://thecjlr.online/index.php/1/issue/view/1

An Evaluation of Research Methods Used in Computer Science Education Studies

Muhammad Babar Hussain
Ms Scholar Department of Computer Science the Bahira University
Abaid Hussain Shah
Ms Scholar Department of Computer Science the Bahira University

Abstract

Methodological evaluations have been effective in uncovering patterns in research and enhancing research procedures in various academic fields. Three methodological reviews have been conducted on the developing subject of computer science education research. However, these evaluations were shown to have deficiencies in terms of their dependability and generalizability. This article provides a thorough, trustworthy, and practical summary of the most recent studies on computer science teaching. This evaluation is important because it provides an opportunity to enhance computer science education practice and tackles the limitations of past methodological reviews. The main focus of our research was to examine the methodological aspects of studies published in reputable computer science education research forums between 2000 and 2005. This investigation included nine specific subquestions. The main purpose of this methodological review is to promote informed discussion about the implementation of computer science education and to establish a strong methodological foundation for recommendations that will further research in computer science education. A sample of 352 articles was randomly selected from a total of 1306 computer science education papers published between 2000 and 2005. The sample was chosen in a way that ensures it is representative of the entire population. Coding was conducted on many aspects of the 352 articles, including their general features, report components, study methodologies, research design, independent and dependent variables, as well as mediating/moderating variables analyzed. Statistical processes were also examined. A second assessor coded a total of 53 items for the reliability subsample. When analyzing the results of the study, suggestions are made to enhance the existing body of research on computer science education.

Keywords- Research Methods, Computer Science Education Studies

Introduction

First page of Fincher and Petre's (2004) "Computer science education research is a developing area that consistently produces a collection of written works" The discipline considers this work to be a seminal work of literature. Notable researchers in the field of computer science education, including as According to Mark Guzdial and Vicki Almstrum, computer science education research needs to integrate methodologies from behavioural research into computer science education and educational research generally if it is to be acknowledged as a new field (Almstrum, Hazzan, Guzdial, & Petre, 2005). By "behavioural research" in this methodological review, we mean the study on learning sciences, education, and cognitive science that Guzdzial alludes to on page 192 of Almastrum et al.'s 2005 publication. In Almstrum

and colleagues (2005), Guzdial brought attention to the fact that there is no link to behavioural research, highlighting that the main challenge in computer education is to avoid duplicating current initiatives. It may be necessary to implement a new approach to education in order to adequately teach and understand the complexities of computers, considering their revolutionary character. The claim is totally false; for the past fifty years, there has been zero change to the underlying mechanics of human learning. Unfortunately, there is a large amount of work in the fields of education, cognitive science, and learning sciences that came before our own work, and this is something that we in the computing education research community tend to overlook. Page 191 and page 192 are the chosen ones. To bridge the gap in understanding between computer science education and rigorous behavioural research, we can examine the present approaches used in the field and compare them to well-established concepts and practices in the field. This research takes a look at the current approaches to computer science education and offers suggestions for how to apply behavioural science principles to enhance computer science education research. Our major goal is to make strides in computer science education research by improving practice, impacting policy, and using other approaches. We want to make this subject go from new to well-established. Computing is the systematic study of information description and modification processes including theory, analysis, design, efficiency, implementation, and application (p. 12), as stated by Delsting et al. (1989).

This review can be useful for many different things. Researchers in the field of computer science education can benefit greatly from learning the standard operating procedures (SOPs) used by their peers. This includes determining the most studied variables, the methods used to measure them, and the methodology used to process and display the results. Additionally, suggestions for how students can improve their own investigation will be provided. Funders, practitioners, and educational administrators who utilise computer science education research will gain a better understanding of the subject's current body of knowledge, including its strengths and shortcomings. Afterwards, decisions on policy and practice might be guided by this understanding. In conclusion, this analysis emphasises the crucial role that important individuals, particularly those who have provided funding, edited, or reviewed research on computer science education, have played in this field. These people decide what kinds of research designs and publication styles are considered acceptable. Before diving into the current analysis, this part provides a brief overview of three previous evaluations of computer science education studies. The reader is invited to delve deeper into the methodologies used in the following sections of this study. Here you can find a concise overview of the following topics: bias analysis, coding book development, sampling method and sample frame, interrater training protocols, and data analytics. Research on computer science education is explored in the results section, which offers a wealth of descriptive statistics. You can learn about the most prolific writers and the most often used statistical analyses from these statistics. Throughout the course of the meeting, we go over the research questions once again.

In conclusion, this paper provides multiple recommendations to improve the methodology.

Literature Review

In order to find any prior surveys on computer science education that could provide useful information for our study, we conducted a literature search. We searched reputable academic databases, such the ACM digital library, multiple times to locate those reviews. We also looked at the contents tables of well-known journals, like the computer science education-focused SIGCSE Bulletin. In addition, we searched the references of all the publications that were relevant. Past work by Valentine (2004), Randolph (in press), and Randolph, Bednarik, and Myller (2005) has evaluated the methods used in computer science education studies that involved secondary or postsecondary students. The following is a condensed version of the reviews.

A Review of K-12 Computer Science Education Program Evaluations

Computer science education program evaluation results were thoroughly reviewed and analyzed by Randolph (in press). (In light of the difficulties in drawing clear lines between research and evaluation, we shall refer to any document that the author has designated as an evaluation, evaluation report, or program evaluation report within the context of this methodological evaluation as an evaluation report.) Computer science education programs for students in grades K-12 were the subject of 29 program assessment papers that Randolph found (in press). This was accomplished by an exhaustive electronic and manual search of credible scholarly literature.

A summary of Randolph's (in press) principal findings appears below:

- The majority of the programs that were evaluated offered direct instruction in computer science for general education courses to secondary school students in North America.
- In descending order of frequency, the evaluators utilized questionnaires, pre-existing data sources, standardized tests, assessments developed by teachers or researchers, stakeholder attitudes, program enrollment, academic achievement in core courses, and computer science achievement as the most frequently employed metrics.

. A single computer science achievement test, which had been assessed for validity and reliability, was discontinued.

- The two most widely used study designs were the pretest-posttest design with a control group and the one-group posttest-only design. A correlation between program type and enhanced accomplishment in computer science could not be identified.

A Methodological Review of Selected Articles in SIGCSE Technical Symposium Proceedings

Valentine (2004) conducted a thorough examination of conference proceedings related to computer science education, covering a period of more than twenty years. The investigation largely focused on the computer science courses completed by first-year students. Valentine scrutinized a total of 444 articles, which were classified into

six distinct groups. According to the analysis, experimental works accounted for 21% of all publications over the past two decades. The author aims to evaluate a treatment in an experimental piece using scientific analysis (p. 256). Valentine's further information can be summarized as follows:

- The proportion of experimental papers has been increasing steadily since the mid-1990s.
- The proportion of papers meeting the criteria for "Marco Polo" (i.e., papers based on personal observation) has been consistently declining in a linear manner since 1984.
- The total number of papers published in the SIGCSE forum has been steadily increasing since 1984. (According to Randolph, Bednarik, & Myller, 2005, p. 104.)

Valentine determined that the problem in computer science education research is that there are too many experimental studies and not enough publications that focus on self-promotion, sharing personal insights, and providing detailed explanations of tools. Due to the absence of inter-rater agreement computations and the only involvement of Valentine as the programmer in the project, the obtained findings are inaccurate.

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

Randolph, Bednarik, and Myller (2005) thoroughly and methodically evaluated every full paper presented at the Koli Calling: Finnish/Baltic Sea Conference on Computer Science Education, often known as the Koli Proceedings, from 2001 to 2004. The approach of each publication was evaluated based on many factors: (a) the arrangement of parts related to methodologies, literature review, and program description; (b) the organization of the reports; and (d) the geographical origin of the publications. Their conclusions were reached after a thorough analysis of every manuscript published in the Koli Proceedings over a period of four years.

- The Koli Proceedings showcased the highest quantity of program (project) description papers.

The majority of empirical articles that cover investigations using human participants mostly utilized survey research and quasi-experimental procedures as research approaches.

- The traditional frameworks found in academic papers on behavioral science differ greatly from those found in empirical studies that study research using human participants. For instance, barely 50% of the papers that described research involving human subjects included literature reviews, and out of those publications, only 17% directly addressed study issues.
- The literature reviews in empirical articles were given little importance and mostly focused on describing the program evaluation.
- The Koli Calling sessions focused on presenting the majority of research on computer science education in the Nordic and Baltic nations, with a particular emphasis on Finland. In addition, none of the papers discussed the accuracy and reliability of the measures, which was an additional finding.

Based on the research conducted by Randolph, Bednarik, and Myller (2005) and

Valentine (2004), there were only a small number of studies that investigated computer science education beyond just explaining how programs work. In the past, impact analysis was often conducted using data obtained from anecdotal instances or using inadequate study designs.

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

Despite the fact that there were three methodological evaluations of research in computer science education, the reviews were not comprehensive, credible, or in-depth. Only a small percentage of the most recent and highly acknowledged research in this subject was covered by the three methodological assessments of computer science education that were conducted in the past (Randolph, in press; Randolph, Bednarik, & Myller, 2005; Valentine, 2004). In addition, the review that Valentine wrote, which is widely regarded as the most authoritative publication in the subject of computer science education research, reveals an insufficient level of trustworthiness and applicability: In spite of the fact that Valentine read a substantial number of articles, he only included those that were pertinent to research in computer science education and were published in a particular forum. Articles that were deemed irrelevant and had no connection to the fundamentals of computer science were not included. Valentine assigned a one-of-a-kind variable to every article and coded each piece on its own, without taking into account the opinions of any other evaluators. In his sole capacity, he was responsible for classifying the pieces into the six categories that are as follows: experimental, John Henry, tools, nifty, philosophy, and Marco Polo. Because it encompasses such a wide range of topics, the experimental category is not appropriate for making recommendations for changes to practice. According to Valentine (2004), operationalization is defined as the process of doing a scientific analysis and evaluation of a "treatment" (p. 256 or 256).

Purpose and Research Questions

For the purpose of determining the generalizability and trustworthiness of a representative sample of research that have been published in significant computer science education forums over the course of the past six years, a thorough review was carried out. For the purpose of addressing the limitations that have been found in previous methodological evaluations of computer science education research, this review was purposefully developed to be both reproducible and reliable. The current methodological review has three distinct advantages over more recent reviews, which are as follows: (a) It offers a significantly more extensive coverage of the field of computer science education; (b) It analyzes articles that incorporate a deeper level of analysis (as well as a more detailed coding sheet) than any previous review; and (c) It demonstrates a higher level of reliability and replicability compared to any previous review. In conclusion, this rigorous evaluation enhances the scope, dependability, and complexity of the previous reviews while simultaneously improving the overall quality of the reviews. In order to facilitate informed discourse regarding the practical application of computer science education research and to advance more severe

recommendations regarding the upgrading of computer science education research, the methodological evaluation served as a key basis for the advancement of these proposals. In the event that our recommendations are implemented and communication is broadened, it is envisaged that the development of computer science education will contribute to the resolution of the social and economic challenges that will emerge in a future that is characterized by a high level of technological innovation. The primary focus of our research was to examine the methodological qualities present in publications published in prominent computer science education research forums between 2000 and 2005.

This is a summary of the primary research topic, which is presented in the following list of sub-questions:

1. What proportion of papers did not reveal experiments involving human subjects?
2. Which categories of papers were published in terms of the proportion of articles that did not report research involving human participants?
3. What proportion of papers that reported research involving human participants relied exclusively on anecdotal evidence to support their claims?
4. What percentage of the articles that reported research involving human participants utilized which methodologies?
5. In the papers that reported research with human subjects, please include details about the specific measures used, the proportions employed, and whether any psychometric information was published.
6. In the papers that reported studies involving human participants, include details about the characteristics and proportions of the elements that were analyzed as mediators, moderators, independent variables, and dependent variables.
7. What were the proportions and types of designs used in the papers that applied experimental and quasi-experimental methodologies? Furthermore, were the volunteers chosen or allocated randomly?
8. Among the papers that offered quantitative findings, what statistical approaches were used and to what extent were they implemented?
9. What were the structural qualities included in the publications documenting research with human participants?

Biases

Both the primary author and the lead researcher have substantial understanding in the subject of behavioral science, notably in the areas of evaluating educational programs and performing quantitative research in the education sector. There is a similarity between the origins of the second and fourth authors. The third author maintains a career in the field of computer science. The primary focus of our research was to investigate the biases that are associated with behavioral scientists who have received training in quantitative approaches. When conducting educational studies that involve human participants, it is essential to adhere to the standards, protocols, and guidelines that are associated with behavioral research. Nevertheless, we recognize that the field of computer science education and research is a dynamic and diverse one,

and that the behavioral science perspective is merely one of numerous viable approaches to examining research in the field of computer science education.

Method

With the help of Neuendorf's Integrative Model of Content Analysis (2002), the technique review was conducted. Here are the steps in the sequential order that make up Neuendorf's paradigm: Here are the steps involved in the method: theory and rationale formulation, variable and measure operationalization, coding form and book creation, sample selection, training, pilot and final reliability calculations, data coding and analysis, and results reporting. We detail our strategy for carrying out each step of Neuendorf's notion in detail in the sections that follow. We shall skip over the logic, the initial step in Neuendorf's paradigm, because we've already covered that.

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

The variables, measures, coding forms, and books were all developed and established before this methodological review. This review is the sixth in a series of reviews that we have conducted previously. For more information, refer to the following sources: Randolph, 2007b; Randolph, in press; Randolph, Bednarik, & Myller, 2005; Randolph, Bednarik, Silander, et al., 2005; Randolph & Hartikainen, 2005; and Randolph, Hartikainen, 2004.

Sampling

Between 2000 and 2005, a proportional stratified random sample of 352 papers was selected from eight prominent peer-reviewed publications on computer science education, without replacement. The sample size of 352 was determined using The Sample Planning Wizard (2005) based on a limited population of 1,306. Resampling was subsequently used to validate the estimation. The sample was categorized based on the publication's year and source. Table 1 displays the number of participants and populations classified by year and forum. The inventory of the 352 articles included in this sample may be found in Appendix A of Randolph's publication from 2007. Within the existing body of research on computer science education, the population was characterized as a construct using the term "population." Certain forums, such as the Journal of Information Systems, which did not primarily concentrate on the education of computer science were not included in the discussion. In the course of the research, the research methodology that has been established in the body of computer science education research was not given specific consideration. As a consequence of this, the research excluded portions of the literature that were less significant and ambiguous. These parts included unpublished reports, program assessment reports, and other publications that had not been subjected to peer review. More specifically, we concentrated on research methodologies that are currently being published in reputable publications that are devoted to research in the field of computer science education and are subject to peer review. These publications come from a variety of conferences and journals, including the SIGCSE Technical Symposium, the Innovation and Technology in Computer Science Education Conference, the Australasian

Computing Education Conference, the Koli Calling: Finnish/Baltic Sea Conference on Computer Science Education, the International Computer Science Education Research Workshop, and the June and December issues of the SIGCSE Bulletin. These publications are included in this anthology. The proceedings of the SIGCSE and ITiCSE conferences are published in the Bulletin volumes that are released in the autumn and spring respectively. After giving it some thought, it has become abundantly clear that the Journal of Information Technology Education ought should have been a part of the sample, despite the fact that we were not aware of this fact at the time. We excluded editorials, book reviews, conference reviews, poster summaries, prefaces, and introductions from the sampling frame, but we did include "full-length papers" in the sampling frame. The selection of samples was typically done by excluding publications that had not been peer-reviewed, as well as those that were condensed into posters and had a length of two pages or fewer. Only peer-reviewed publications were considered for inclusion in the Bulletin's sampling frame; reports from working groups, invited columns, and featured columns were not. Both the CSE and the JCSE did not permit introductions or editorials. Articles have to be at least three pages long and pass a strict peer review process in order to be considered for the SIGCSE, ITICSE, ACE, and ICER forums. We did not permit panel discussions or short papers (those having two pages or fewer). Only research and discussion papers were authorized to be presented at the Koli conference; no poster or demonstration presentations were.

Training and Determining Pilot Reliabilities

Appendices B and C of Randolph's publication (2007a) provide the coding sheet and book provided to an experienced individual assigned as an interrater reliability reviewer for conducting methodological evaluations. The interrater reliability reviewer and the first author collaborated to examine the code sheet and the book, addressing any inquiries that arose during the process. All discrepancies or inquiries identified in the code book or coding page were addressed in light of the concerns raised during the initial training session. Afterwards, the coding book and document were modified and dispatched to the autonomous reliability reviewer. A deliberate experimental sample of ten research articles on computer science education was utilized, and the participants were instructed to autonomously encode them. Please note that some articles were excluded from the final dependability subsample. The purposive sample consisted of anecdotal stories, non-participant articles, and tales that the original author believed represented the several research methodologies being considered. The individuals responsible for coding those ten parts were the original author and lead programmer. After completing the ten articles, both coders convened to share their observations, address any uncertainties regarding the coding page or book, and rectify any errors or ambiguities. If a disagreement arises among the coders regarding article codes, the original author will revise the coding guide after investigating the reason of the problem. The ultimate reliability subsample was encoded subsequent to pilot

testing and subsequent adjustment of the coding manual and coding (for further information, refer to the section titled Calculating Final Reliabilities).

Coding

To learn all about the coding variables, where they came from, and how the coding process worked, check out the coding document and book in Randolph's (2007a) Appendices B and C. There is a lack of detail regarding the comprehensive coding document and coding book in the study. Randolph (2007a) does, however, include these materials as appendices. Over 120 variables were encoded in all. Categories for the aforementioned features include demographics, article type, methodology, research design, dependent and mediating measures, examined independent variables, and other relevant criteria.

Calculating Final Reliabilities

According to Neuendorf (2002), determining the degree of agreement amongst raters requires only a subsample of 50-200 units. In order to determine the articles' credibility, 53 were chosen at random from a pool of 352 articles. The 53 articles were coded independently by the reviewer specializing in interrater reliability in order to evaluate their dependabilities. Brennan and Prediger (1981) advised using the free-marginal kappa (κm) as an interrater agreement measure because the amounts of each level of variables to be coded were not predefined. (When we say that there was no room for maneuver in the distribution, what we really mean is that there was no predetermined criteria for how many things should go into each category.) Look at the 1981 book by Prediger and Brennan. Kappa levels below 0.4 were considered undesirable, values between 0.4 and 0.6 were considered poor, values between 0.6 and 0.8 were considered fair, and values above 0.8 were considered good when evaluating the reliabilities. Through the process of resampling, confidence intervals that incorporate kappa were computed. In order to frame the kappa statistic, Randolph (2007a) supplies the confidence intervals.

Data Analysis

To address the primary aim of the research, frequencies were calculated for each set of binomial or multinomial variables. Resampling was employed to produce the 95% confidence intervals for every multinomial category or binary variable (Good, 2001; Simon, 1997). One reason for the increasing popularity of this alternative inductive method to significance assessment is the difficulty in applying classic significance tests to complicated samples (Garson, 2006, n.p.). The Resampling Stats software, released in 1999, was utilized alongside Grosberg's resampling program, which does not have a specified date of publication. Randolph (2007a) includes code snippets in Appendix E that can be utilized to calculate confidence intervals for a given percentage using Resampling Stats.

Results

Interrater Reliability

Neuendorf (2002) found that the interrater reliabilities of the researchers were mostly excellent or sufficient. However, the reliabilities of the following six variables were less

than 0.60: The criteria established by Kinnunen (n.d.) include: the paper's kind, the inclusion of human participants, the presence of a literature review, the level of depth in describing the setting and method, and the adequacy of separating the results and commentary. The reliability of five out of seven factors pertaining to report items was found to be low. To calculate the κm, which is a metric used to assess inter-rater reliability, Randolph (2007a) provides information on the number of cases (out of 53) that can be utilized, along with the corresponding 95% confidence intervals.

Discussion

Study Limitations

One constraint of the study was the limited agreement among raters on a small portion of the variables. In order to overcome this limitation of the study, we implemented qualifiers on claims supported by variables that had low reliability or abstained from drawing definitive conclusions about these variables. As shown in the Methods section, our focus mostly revolved around the perspective of behavior scientists with a quantitative orientation. Due to our greater interest in quantitative experimental research, we did not fully prioritize publications that simply relied on qualitative techniques of inquiry. Due to the diverse and unpredictable nature of qualitative approaches, we had reservations about our ability to create and execute a dependable system for categorizing, assessing, and examining those documents. Therefore, this research has yet another constraint.

Revisiting Research Questions

The primary inquiry we posed was "What are the methodological attributes of studies recorded in articles published in prominent computer science education research forums from 2000 to 2005?" We devised nine sub-questions to address this topic. Below are abridged versions of the responses provided to the research inquiries.

What is the number of publications that did not involve any experiments with human subjects?

Approximately one-third of the publications were discovered to have no studies involving human subjects. The aforementioned articles consisted of program descriptions, theoretical or methodological pieces, and reviews of related literature. The current review's proportion (33.8%) is approximately 30% lower than the review conducted by Randolph, Bednarik, and Myller (2005) on papers published in the proceedings of the Koli Calling conference.

Which types of publications were released in areas where human beings were not involved, and what proportion of those papers belonged to each category?

Among the studies that did not involve human subjects research, 60% solely consisted of treatments descriptions without any evaluation of their effectiveness for computer science students. The fraction of program descriptions quantified in previous computing methodology evaluations is similar to or somewhat greater than the proportion of publications in question. The program description category we have is highly comparable to the Marco Polo and Tools categories in Valentine's (2004) study. Therefore, it is probable that the conclusions from Valentine's research can be

applicable to our field as well. Valentine's (2004) study reveals that around 50% of scientific articles in computer science education belong to this group. Our categorization method found that 43% of the computer science publications examined by Tichy, Lukowicz, Prechelt, and Heinz (1995) were articles related to design and modeling, or program descriptions.

What is the number of articles that relied solely on anecdotal evidence to support their assertions in extensive investigations involving human subjects?

Efforts to tackle the issue of excessive reliance on anecdotal evidence in computer science research, namely in the field of software engineering, have been ongoing for over a decade. "According to Holloway (1995, p. 21), empirical claims about software engineering are seldom supported by logical or empirical evidence." "It is absurd, yet widespread, to base an entire discipline on such an unstable epistemological foundation." Table 9 indicates that a further issue with the current state of research on computer science education is the abundance of anecdotal information. It is crucial to bear in mind that in this review, "anecdotal evidence" pertains to a researcher's casual observations of a phenomenon. We do not imply that individuals possess an innate incapacity to make precise and reliable observations. Such a notion would contradict ethnographic studies and other forms of study in which individuals examine behavior in a practical and objective manner, while also operationalizing it. Moreover, we are in mutual agreement that anecdotal evidence plays a crucial role in assisting researchers in developing ideas. Conversely, Holloway (1995) highlights the numerous issues associated with relying on unofficial anecdotal material to substantiate concepts. Valentine (2004) found comparable results in his methodological examination about the prevalence of anecdotal evidence in computer science education studies. The importance of collecting empirical data has been highlighted in various computer science education research initiatives, such as those carried out by Holmboe, McIver, and George (2001) and Clancy, Stasko, Guzdzial, Fincher, and Dale (2001).

What were the specific methodologies utilized and how frequently were they referenced in the papers that described research involving human subjects?

Approximately 66% of the computer science education papers used experimental research in the evaluation of this article (refer to Table 11). In the following section, we will demonstrate that the prevalent experimental designs were susceptible to nearly all potential sources of internal validity threats. Based on the data, experimental procedures were more prevalent than qualitative methods. Researchers have the option to employ qualitative approaches, experimental procedures, or quasi-experimental procedures, or a combination of these, in order to establish causal relationships and evaluate their hypotheses (Mohr, 1999). The fundamental basis of experimental and quasi-experimental research, as described by Mohr, is empirical causal reasoning. This involves comparing an actual situation with a hypothetical state that did not occur. The core of qualitative research is physical causal reasoning, often known as the Modus Operandi Method as stated by Scriven (1976). The majority of research conducted on computer science education employ methodologies that have

the potential to establish causal relationships, given the appropriate conditions. This is excellent news for the industry.

Did the articles documenting research on human participants provide any psychometric data, such as measurements or quantities?

Undoubtedly, surveys were the most widely used method for collecting data. Table 17 indicates that somewhat more than 50% of the evaluations were conducted using questionnaires. Regarding frequency of usage, educator- or researcher-generated assessments ranked third, with grades following closely behind. Wilkinson and the Task Force on Statistical Inference urge that a researcher should provide a concise summary of the psychometric qualities of the scores derived from a questionnaire, while considering how the instrument is used within a specific community. It is disconcerting that among the 65 research that utilized questionnaires, only one included details regarding the instrument's validity or reliability. The qualities of a psychometric tool encompass its reliability, validity, and internal validity (1999, n.p.)... A significant deficiency in the literature on computer science education is the lack of psychometric data pertaining to the instruments.

Which kinds of mediating, moderating, independent, and dependent factors were studied, and in what proportions, by the publications documenting research involving human participants?

Mark Guzdzial, a member of the Challenges to Computer Science Education Research working group, states that student opinions are not reliable indicators of the quality of instruction or learning, as widely known (Almstrum et al., 2005, p. 191). However, based on this data, it is evident that attitudes are the most frequently subjected to testing. Opinions constituted the sole independent component in 44% of the articles. Guzdzial has highlighted that attitudes may capture the interest of researchers in computer science education, but they should not be seen as dependable indicators of student learning or instructor performance.

The posttest-only with control and one-group posttest-only designs were the most widely used experimental study methodologies. The design that ranked second in popularity, which incorporated controls, had an implementation rate that was more than twice as high as that of the one-group posttest-only design. If the objective was to establish causal inference, it would have been more advantageous to employ various designs involving pretests and/or control groups. Shadish, Cook, and Campbell (2002) assert that nearly all internal validity concerns can impact a one-group post-test design. Our examination of the articles' selection and assignment methods indicated that 87% of the students voluntarily took part in the study by selecting either the treatment or control groups. An astonishing 86% of the published papers utilized convenience samples. Several scholars, including Kish (1987) and Lavori, Louis, Bailar, and Polansky (1986), have investigated the formal model of sampling, which involves first obtaining a random sample and then randomly assigning individuals to different groups. Certain organizations contend that formal sampling approaches lack utility, as demonstrated by Shadish et al. (2002) and Wilkinson et al. (1999) of the Task Force

on Statistical Inference. The debate is currently in progress. Academic research on computer science education suggests that deliberate sampling is more favorable than random selection, where it is feasible to implement. Purposeful sampling entails assessing the representativeness of a sample by identifying apparent similarities, excluding extraneous data, and distinguishing between various samples. Alternatively, one can explore potential causal explanations and using techniques like as interpolation and extrapolation. When it comes to randomly allocating individuals to treatment conditions, the same principles are applicable. When random assignment is not possible, there are alternative methods to establish causation, but random assignment is typically the preferable approach. Confounding variables, defined as variables that impact the observed connections between a causal variable and an outcome, must be assessed and removed from experimental designs or analyses when it is not feasible to randomly assign participants to different conditions. An alternative strategy would involve inquiring about the participants' computer proficiency and utilizing this data to statistically adjust for their level of expertise.

Within the papers that presented quantitative findings, what was the total number and specific types of statistical methodologies employed?

When conducting certain statistical studies, the American Psychological Association advises including specific information as outlined in their publication from 2001, page 23. A set of sufficient statistics in the context of parametric tests of location includes cell means, cell sample sizes, and measures of variability. Alternatively, it can consist of cell means, mean square error, and degrees of freedom associated with the effect being tested. Furthermore, it is advisable to include an effect size in the publication along with p-values, as suggested by the American Psychological Association (2001) and the American Psychological Association's Task Force on Statistical Inference Testing (Wilkinson & Task Force on Statistical Inference, 1999). Quantitative data are presented with inferential analyses being conducted in 36% of cases, according to the findings of this review. A significant proportion of computer science teachers who perform inferential analyses are able to provide informative statistics. Data characterization, including nonparametric methods, correlation or covariance matrices, and measures of central tendency and dispersion for parametric studies, are all crucial components of any correlational investigation.

What specific architectural elements were mentioned in the articles that described experiments involving human participants?

Several intriguing findings were made about the contents of the analyzed publications. Empirical studies often lack literature reviews (about 25% of the time), study questions (around 22%), and sufficient details regarding the instruments or procedures employed (less than 50% of the time). Nevertheless, it is uncertain if the papers included in this review can be classified as literature reviews due to the lack of consensus among reviewers regarding the reporting of some characteristics. Additional clarification is required for critical components of the report, such as the literature review, research questions, and participant descriptions, as two raters are unable to

reach a consensus on their presence. When two raters have conflicting opinions regarding the presence of a literature review in an academic journal, there is an increased probability that the review is flawed. The ACM SIGCSE Working Group on Challenges to Computer Science Education found that computer science professors frequently duplicate efforts and that there is insufficient evidence to support this claim (Almstrum et al., 2005, p. 191). If this assumption holds true, it implies that research articles on computer science education do not incorporate literature reviews. Mark, Henry, and Julnes (2000) argue that strong prior evidence often lowers the threshold for subsequent evidence. In other words, researchers performing comprehensive literature studies do not need to collect as much evidence to support their findings. In addition to the need for accumulating evidence, it is important to avoid duplicating efforts that have already been undertaken. Furthermore, the report's dubious literature review and other sections give the impression that the majority of studies on computer science education do not adhere to the format recommended by the American Psychological Association. There is a notable discrepancy between these two reporting customs. Consensus structures provide authors the opportunity to communicate their discoveries through many means and promote diverse reporting approaches, although they present a challenge for readers in extracting the desired information from the articles. Furthermore, it is hypothesized that the absence of established guidelines for computer science education papers may result in novice researchers omitting crucial information, such as the methodology employed and the participants included in the study. The report component factors, such as insufficient information about participants or procedures or a lack of a complete literature review, were only found in articles discussing research with human beings. However, these criteria were not relevant as the report structures of theoretical publications and program descriptions differed greatly from those of studies involving human subjects.

Conclusion

We carried out a methodical examination of the papers that were published in respectable computer science education forums between 2000 and 2005 using a content analysis approach. From all the papers published at that time, 352 articles were chosen at random to form a sample. Every article underwent a thorough examination of its general features, techniques used, research design, independent and dependent variables, mediating or moderating variables, measures used, and statistical procedures followed. The review's major findings are enumerated here:

. Any research involving human beings was absent from one-third of the publications. The bulk of the studies without any mention of human participants were program descriptions.

● Almost forty percent of articles involving human participants relied only on anecdotal evidence to back up their assertions.

● Qualitative, quasi-experimental, or experimental techniques were used by a sizable percentage of the publications that offered more information than anecdotal evidence.

● Most studies using an experimental research approach used a single group posttest design.
● The most common independent, mediating/moderating, and dependent factors, in that order, were found to be student instruction, gender, and attitudes.
● Questionnaires were clearly the most often utilized kind of measuring tool.
The great bulk of measurement equipment that could have used psychometric information had it conspicuously missing. Inferential statistics was also often used on insufficient statistical data.

Computer Science Education Research at the Crossroads

The review's conclusions indicate that computer science educators have developed a substantial amount of well-informed research proposals through ad hoc investigations or anecdotal evidence. However, they have not conducted a comprehensive examination of these theories. The computer science training is currently discontinued. Computer science education scholars at the intersections bring numerous legitimate assumptions to the table. Currently, the outcome of these theories remains uncertain. These extensively researched theories, which have repeatedly been shown false through the examination of "success stories" and persuasive sales presentations (Holloway, 1995, p. 20), may eventually be universally acknowledged as truths. However, there is no actual evidence to support this claim. They will be encompassed by conventional wisdom, to put it simply. Because there is no empirical research proving their validity, the findings are referred to as folk conclusions instead of folk theorems (see to Harel, 1980) or folk beliefs (refer to Denning, 1980). Given the progressive accumulation of scientific information over time, incorporating well-supported theories into popular conclusions would inevitably include both accurate and inaccurate popular beliefs in the field of computer science education. The study of computer science has the potential to evolve into a scientific discipline over time, since it builds upon untested assumptions to expand its knowledge base. Holloway (1995, p. 21) argues that it is absurd to base an entire field of study on a weak foundation of knowledge. Equally absurd is wagering on the outcome of an entire discipline relying on dubious epistemology. Nevertheless, we maintain that research endeavors, such as the formulation of hypotheses, should not be completely eliminated from computer science curricula. Having access to a diverse array of approaches is crucial for facilitating a broad spectrum of research activities. Innovation and hypothesis creation are closely interconnected. We argue that the current challenges and concerns in computer science education should determine the extent to which research methodologies are employed. If the current challenges are attributed to insufficiently gathered evidence and lack of rigor, as determined by the Working Group on Challenges to Computer Science Education of the ACM SIGCSE, it would be logical to shift away from relying on anecdotal evidence and hypothesis generation, and instead employ rigorous methods to validate hypotheses (Almstrum et al., 2005). Returning to our original point: achieving a viable future for computer science

education necessitates building upon current theories while also striking a balance between innovative concepts and proven approaches.

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). Publication 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.