Methodological Rigor and Theoretical Foundations of CS Education Research

Alex Lishinski
Michigan State University
College of Education
East Lansing, MI
lishinsk@msu.edu

Jon Good
Michigan State University
College of Education
East Lansing, MI
goodjona@msu.edu

Phil Sands
Michigan State University
College of Education
East Lansing, MI
Phil.sands@gmail.com

Aman Yadav
Michigan State University
College of Education
East Lansing, MI
ayadav@msu.edu

ABSTRACT

The problem of the lack of rigor in CS education research has frequently been discussed and examined. Previous reviews of the literature have examined rigor on both theoretical and methodological dimensions, among others. These reviews have also looked at differences in indicators of rigor between conference proceedings and journal publications. However, to date there is no comprehensive review that has examined the intersection of methodological and theoretical quality.

This paper reports results from a literature review in which we analyzed both the use of theory and methodological rigor of four years of CS education research from the Computer Science Education (CSE) journal and the proceedings of the International Computing Education Research (ICER) conference. The goal was to provide an updated and expanded picture of the methodological quality and use of theory in the most rigorous CS education publications, as well as to compare between conference proceedings and journal publications on these dimensions. Our focus was on research that draws upon learning theory from education, psychology and other disciplines outside CS education.

The results of our review show a different picture than earlier reviews. Focus on empirical results in conference proceedings articles has surpassed that of journal publications, and empirical studies are significantly more likely to make use of theory from outside CS education. Overall, our analysis shows a significant increase in the proportion of articles drawing on theory from outside CS education, compared to earlier literature reviews, whereas indicators of methodological quality show no such change.

ICER '16, September 08 - 12, 2016, Melbourne, VIC, Australia

ISBN 978-1-4503-4449-4/16/09...$15.00

DOI: http://dx.doi.org/10.1145/2960310.2960328

Keywords

computing education, research methods, literature review, theoretical grounding

1. INTRODUCTION

Computer science education is an education field, but by and large it does not have the same mature and independent status of other content area education fields. Computer science education research is largely done by practitioners, computer scientists who teach and computer science teachers, rather than dedicated education researchers with specialized training in social science research methods and theory [17]. For this reason, there is much concern about the rigor of computer science education research. The field, although decades old, can be considered young in terms of scholarly development [17]. As a practitioner driven field, rather than a researcher driven field, there are a few major areas for improvement in CS education research.

The rigor of CS education research has been previously examined on a number of dimensions. The degree to which research relies on empirical results [1], the methodological rigor of those empirical studies [16], the diversity of research methods employed [2], and the degree to which relevant theory from the behavioral sciences and other education fields is used to ground and guide research [6]. With respect to the last of these dimensions, Mark Guzdial made the statement at a 2005 SIGCSE panel that "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" [2]. In an editorial for a 2015 special issue of Computer Science Education, Anthony Robins quoted Guzdial and argued that his assessment of CS education research remained true 10 years later.

Fincher and Petre (2004) gave six broad principles for CS education research. They are:

- Pose significant questions that can be answered empirically
- Link research to relevant theory
- Use methods that permit direct investigation of the question

- Provide a coherent and explicit chain of reasoning
- Replicate and generalize across studies
- Disclose research to encourage professional scrutiny and critique

The present study sought to investigate the degree to which CS education research addresses the first three of these principles. To that end, we present the details of a literature review of recent publications in *Computer Science Education* (CSE) and the *Proceedings of the International Computing Education Research conference* (ICER) in which we sought to answer the following research questions:

- To what extent do CS education research articles make use of previous theory from outside CS, and how does this compare to the results of previous reviews?
- To what extent do CS education research articles present empirical results, what level of methodological rigor is used, and how do these results compare to those of previous reviews?
- To what extent do CS education research studies make explicit their research questions, and how does this compare to the results of previous reviews?
- Is there any connection between the methodological and theoretical quality of research?

2. THE ROLE OF THEORY IN EDUCATION RESEARCH

Theory plays an essential role in education research. Suppes argued that education research as a field initially grew as a result of embracing empiricism, but later faltered as a result of inattention to theory [19]. Theory provides the deeper and enduring foundation for the importance and significance of education research [19]. Suppes argued that the importance of theory to education research can be seen by analogy to the role played by theory in the successes of the hard sciences and other social sciences like psychology and economics. The role of theory in the hard sciences lies in its power to organize experience in more deep and rigorous ways, and only through such rigorous organization of experiences is it possible to address the underlying complexity of the phenomena being studied. This is only more true in the behavioral sciences. Superficial empiricism misses essential elements of any subject of scientific study, and in its most extreme form, provides no generalizable principles whatsoever [19].

In scientific fields, by 'theory' is meant a certain kind of explanation, that offers a 'how' and a 'why' something works the way it does [8]. To build theory is to construct a connected set of ideas that form an organized way of explaining particular phenomena [8]. McMillian and Schumacher argued that theory can be seen as beholden to the following criteria: provide parsimonious explanations of observed phenomena, be consistent with the existing body of knowledge, provide a mechanism for verification and revision, and stimulate further research [11]. Thus conceived, theory protects against unscientific approaches to inquiry [8]. Underlying assumptions are made explicit, predictions and explanations for observed phenomena are provided, and perspectives for understanding some class of phenomena are provided by theory [8]. Furthermore, given the societal importance of the phenomena studied in education and behavioral sciences, particularly the fact that social policy is often informed by behavioral science research. Therefore, Tellings argued, solid, grounded, generalizable theoretical models are needed to guide research in the behavioral sciences [20].

2.1 Theory in Computer Science Education Research

The second principle for CS education research is to link research to appropriate theory. Fincher and Petre characterized CS education as an essentially interdisciplinary field. This characterization comes both from the essential fact that computer science education is rooted in both a behavioral science field and a STEM field, as well as the more contingent fact that computer science education is among the more immature education fields. CS education has traditionally borrowed methods and concepts from other fields, including mathematical content areas like physics, as well as other behavioral sciences like psychology. Fincher and Petre described interdisciplinary fields as a 'trading zone' where disciplines come together and ideas, theories and methods are exchanged [6]. General guidance on the study of learning can come directly from theory, as well as indirectly, by means of analogy to educational research done in other similar disciplines [6]. Computer science education borrows a number of different types of research tools from the broader field of education, such as theories, models, and instruments [6]. Examples of theories include Bruner's work on cognitive development, which emphasized the relationship between cognitive process and how well structured disciplinary content is, Vygotsky's work on the Zone of Proximal Development (ZPD), which conceptualizes limits to the amount of learning that can be expected given the student's current developmental state, and Lave's work on how learning must be conceptualized as inextricably situated in authentic contexts, both concrete and social. Models of educational processes include Bloom's Taxonomy, which conceptualizes assessment in education by reference to a hierarchical set of cognitive behaviors that serve to indicate what has been learned, as well as Kolb's learning cycle, which describes the cyclical process by which concrete experiences are turned into abstract representations [6]. Instruments include the Motivated Strategies for Learning Questionnaire (MSLQ) [13], a set of motivational scales, and the Wechsler Adult Intelligence Scale, an intelligence test.

3. METHODOLOGICAL QUALITY IN CS EDUCATION RESEARCH

As in all research fields, methodological quality is a very significant concern for CS education research. The third principle of CS education research is to use methods appropriate to the research questions being investigated. Fincher and Petre framed the relationship between research methodology and rigor in terms of bias-reduction. All studies should provide evidence for their claims, but the strength of the evidence provided by any given study is a function of bias-reduction, which is achieved through appropriately rigorous methodology, among other things (e.g., operationalization of constructs, interpretation of results) [6]. Fincher and Petre detailed the bias reducing properties of different research

designs and methods. For example, a within-subjects design (pre/post) reduces bias due to individual differences over a between-subjects design (post only) [6]. Likewise, an experimental design reduces bias more so than a quasi-experimental design, because random assignment eliminates selection bias.

Research questions are likewise an important part of CS education research. Fincher and Petre warned against the backwards approach of designing your study to explain an already observed result. Experiments designed without a specified research question can lead to bias in a number of ways, because the research questions guide the use of theory and development of methods and interpretation of results. For this reason, research questions should be identified before specifying the methods of the study [5]. Research questions in quantitative studies make explicit the variables of interest and the relationships between them which are to be tested, which in turn suggests the methods that should be used [4]. Research questions in qualitative studies, on the other hand, should specify the main concept of interest and connect to a particular strategy of qualitative investigation, but the questions should be expected to evolve and be refined through the course of investigation [4]. Research questions differ in the degree to which they fit the theory and methods used, but in this study we limit our analysis to whether or not research questions are specified.

4. PREVIOUS LITERATURE REVIEWS IN CS EDUCATION

Literature reviews in CS education are not uncommon, and they often focus on areas where the literature lacks rigor. One common theme in past reviews has been empiricism. A series of reviews has looked at the degree to which research presented at SIGCSE, the largest CS education conference, presents empirical results. The investigation of lack of empirical rigor has shown a consistently improving trend. A review of SIGCSE studies from 1984-2003 found that 21% of studies were 'experimental' (their definition of this term is explained further below) [21], a similar review of studies from multiple CS education research venues found 40% to be experimental [16], whereas a later review of SIGCSE research from 2014-2015 found over 70% of studies to be empirical [1]. This increasing trend is heartening to those concerned about rigor in CS education research, but it is not without caveats. In the most recent review, although 70% of studies were classified as empirical, the quality of such empirical evidence was questionable. Of the 70% of studies classified as empirical, the most common source of data was student feedback surveys [1].

Randolph et al. looked at methodological quality in CS education research across several publication venues [16]. The focus of this literature review was to compare journal publications with conference proceedings publications with respect to 5 indicators of methodological rigor. The indicators of methodological quality were attitudes-only outcome variables, one group posttest-only designs, experimental designs, use of qualitative methods, and reporting only anecdotal evidence. The authors found no statistically significant differences between journal and proceedings publications on these indicators, suggesting that both types are equally rigorous.

Malmi et al. reviewed the use of theory in computer science education research articles from Computer Science Education, the proceedings of ICER, and Transactions of Computing Education [10]. Using their theories, models, and frameworks (TMF) construct, the authors found that about half of all papers published in these outlets did not use any sort of TMF. Looking just at CSE and ICER papers, there was an equal use of TMFs, with articles from both publications using some TMF 57% of the time. The authors further classified the identified TMFs by reference discipline, finding that TMFs came from CS education research in 16% of cases, Computing in 24% of cases, education in 29% and psychology in 19%. Specific TMFs were also counted, but the authors found such diversity in TMFs (314 instances of 226 distinct TMFs; 200 TMFs mentioned only once) that this analysis was not particularly informative.

Previous literature reviews have also looked at categories of CS education research. Joy et al. examined research across many different outlets for CS education research in order to create a taxonomy of research [7]. The basis for their analysis was the notion of a continuum running from a purely technical perspective to a purely educational perspective, comprising four major categories. These were (a) system, which were studies on educational software focusing on the technical, (b) technology, studies on educational software focusing more on the educational but lacking implementation, (c) practical pedagogy, which were educational studies that report results not connected to substantial educational theory, and (d) theoretical pedagogy, studies that were substantially grounded in education theory. Joy et al.'s study provides a taxonomy of the types of studies that are prevalent in different CS education journals, but what has not been looked at heretofore in reviews of the CS education literature is the degree to which CS education research draws theory from the behavioral sciences themselves and from outside CS by analogy, as well the sorts of theoretical themes that are prevalent in the research, and shifts over time in this composition [7]. Other literature reviews have created similar categorizations of different types of computer science education research [9, 18].

5. FOCUS, NEED AND PURPOSE OF STUDY

Our focus in this study was investigating the rigor of research in computer science education research, on both the methodological and theoretical dimensions. It is these two dimensions that we thought were most relevant to assessing the degree to which CS education research is becoming a more mature education discipline along the lines of mathematics or science education. We chose to focus our attention on articles from ICER and CSE, because the previous studies on rigor in CS education research suggest that among journals, CSE is likely to contain the most rigorous work, and among conference proceedings, ICER is likely to contain the most rigorous work [7].

Our investigation of the methodological rigor of CS education research follows from that of Randolph et al. [16]. We examined articles for methodological quality along the 5 dimensions chosen by the authors in that study. However, we felt that the dimensions examined in Randolph et al. [16] were not extensive enough, so we expanded our framework to contain several additional dimensions.

Our investigation of the theoretical rigor of CS education research is a continuation of the work done by Malmi et al. [10]. However, our focus differed from that work, in that we

chose to focus our investigation on the use of theory drawn specifically from work in psychology, education and other social science disciplines that seek to explain the processes by which learning takes place, to give a rough definition. Our reason for restricting our focus in this manner comes from the arguments made in Fincher and Petre [6] and Almstrum et al. [2]. Fincher and Petre argued that computer science education is an inherently interdisciplinary field, and so we thought that an important measure of how mature the discipline has become is the degree to which it reaches for guiding theory from outside itself [6]. The focus on learning theory is similarly motivated by the point made in Almstrum et al. and Robins, that CS education has tended to not make use of previous research from the learning sciences (broadly construed) [2, 17].

The purpose of our study was to examine the theoretical and methodological dimensions of rigor in the same group of papers. We were interested to find out whether the two dimensions of rigor were related in any quantifiable sense. We also thought that the previous work on methodological rigor [16] could be expanded upon. Our goal was to improve the analysis of methodological rigor by using a more detailed set of indicators. We also shifted focus in our analysis of theoretical rigor compared to previous work [10], focusing on the use of theory from outside computer science.

The need for our study is to both provide updated and improved data on these two dimensions of rigor. Randolph et al. looked at articles up to 2005, whereas Malmi et al. looked at articles up to 2011 [10], so the last four years of articles have not been examined on these dimensions of rigor. Our study provides more current data on the rigor of CS education research, which can be compared with that reported in previous studies, and from which more conclusions about the trajectory of rigor in computer science education research can be drawn [7].

6. METHODOLOGY

6.1 Selection of Articles

Articles reviewed for this study include selections from 4 years (2012-2015) of Computer Science Education (CSE), and the proceedings from ICER. There are other outlets for CS education research that could have been included as well, but these 2 were selected because they cover 2 primary venues where rigorous, theory based empirical articles are likely to be published. ICER is a research conference that places a far greater emphasis on both empiricism and theoretical grounding than other conferences, and CSE is a journal that features work that is both empirical and theoretical in nature. We coded all articles from both publications in the years 2012-2015. The rationale for the range of years used was that the most recent year examined in Malmi et al. was 2011 and we wanted to have continuity and comparability with the results of that study.

6.2 Categories of Methodological Quality

Drawing from Randolph et al.'s study [16] on methodological quality in CS education research, we looked at several methodological dimensions. Randolph et al. used 5 indicators of methodological quality to examine all CS education research studies that had human subjects. They were:

- *Attitudes Only*: These studies measured only students' or teachers' attitudes toward the intervention studied. This is distinct from studies where the sole outcome measure is some form of attitudes that are not related to the intervention, such as attitudes toward CS, or self-beliefs.

- *Anecdotal Only*: These studies reported only subjective accounts from the author about the quality of an intervention or instructional approach. This is distinct from rigorous qualitative research, in which subjective observations are coded and accounted for within some kind of rigorous framework

- *Single Group: Posttest Only*: These studies reported quantitative data, but only for one group (no control group) and only after the intervention (no pretest control). This is distinct from experimental and quasi-experimental designs that use posttest only data to do a multiple group comparison.

- *Experimental Designs*: Randolph's definition of experimental included essentially all studies in which quantitative data was collected [15]. This includes all manners of quantitative research design (true experiment, quasi-experiment, single group) and data collection plans (posttest only, pretest and posttest). We collected data on this category for comparability of results, but we also collected data on studies that were experiments in the technical sense of the word, with randomly assigned treatment and control groups.

- *Qualitative Research*: These studies involved non quantitative data that was rigorously analyzed or coded for themes. The data can include coded interviews with students, coded samples of student work, coded transcripts of classroom discourse, or any other type of data that is not quantitative in nature. This is distinct from anecdotal data in which qualitative observations may be reported, but they are not done in a rigorous, thoroughgoing manner.

We modified the coding framework from Randolph et al. [16]. We coded all articles for these original 5 dimensions so that our results would be comparable, but the framework was adjusted and expanded to capture elements of methodological quality not captured by the original framework to add additional levels of nuance to the original framework beyond the simple good/bad division used by Randolph et al. [16]. The expanded framework was structured as follows:

- *Qualitative Data*: The qualitative data category comprised all studies that reported non-quantitative data. Of these studies, we coded them into two groups, Qualitative and Anecdotal, corresponding to the categories from Randolph et al.'s [16] original framework described above. We coded those studies as qualitative that had a systematic protocol or framework for analysis, and as anecdotal those studies that simply reported subjective results.

- *Quantitative Data*: This quantitative data category comprised all studies with any sort of quantitative data. Of those studies, we coded them into three groups: Experimental, Quasi-experimental, and Single Group. Experimental studies were those which used multiple

groups with completely random assignment. Quasi-experimental studies were those which used multiple groups without completely random assignment. Single Group studies were those which used just a single group for evaluation. All of these categories were given distinct codes.

- *Data Collection Procedure*: Of all quantitative studies, we further characterized them on the basis of their data collection procedure. We coded studies into two groups, posttest only, and pre/post. Posttest only studies collected data only after an intervention, whereas pre/post studies collected data before and after an intervention.

- *Mixed Methods*: If the study in question used both quantitative and qualitative measures, we coded the study as mixed methods in addition to the coding for the quantitative and qualitative methods.

The quantitative studies were also coded for the type of outcomes they measured. The categories were attitudes only, psychological, data mining, overall grade, and other. We coded for the type of outcome measure as a way of expanding on the coding of studies that used an 'attitudes only' outcome. In addition to counting those studies, we wanted to get a sense of what other types of outcomes CS education studies are using, which should be a very rough proxy for the type of conclusions that these studies lend themselves to drawing.

- *Attitudes only* corresponds to the original category from Randolph et al. [16] described above, corresponding to studies in which the only outcome was the attitudes of participants toward the study intervention.

- *Psychological* refers to all broadly psychological measurement instruments, for example, surveys on self-efficacy, surveys on attitudes towards computing, and cognitive tests of psychological constructs like problem solving.

- *Data Mining* refers to outcomes that were the result of an automatic data collection process such as data from compilers, IDEs or LMSs.

- *Overall Grade* refers to outcomes based on overall course grades or final exam grades in the course where the intervention took place.

- *Other* refers to any outcomes not falling in one of the above categories.

These revised categories correspond to more detailed levels of methodological rigor. In addition to these levels, we also decided to code articles for whether or not explicit research questions appear in them. We thought that the inclusion of explicit research questions is also an indicator of methodological rigor, as authors must formulate a substantive question that their results can answer, rather than simply presenting a series of results that, while perhaps significant and interesting, may not substantively add to the theoretical knowledge of the topic. For that reason we coded all articles as a yes/no for whether they included explicit research questions.

6.3 Qualitative Coding for the Use of External Theory from the Learning Sciences

Similar to Malmi et al., we examined all articles as well for their use of theory from the learning sciences, with a particular focus on just studies from outside computer science education. We classified particular studies as making use of outside theory if they contained at least one citation to a reference on learning theory from outside CS education. We considered learning theory studies to be theoretical and/or empirical papers/books from psychology, cognitive science, etc. In other words, anything that derives from some abstract model of how students learn or understand. This can be direct references to psychological theory pieces (e.g. Constructivism in Education), Empirical psychology pieces (e.g. An Empirical Examination of Motivational Issues in Problem Solving in Middle School Students), or articles from other disciplines that explicitly invoke such theory (e.g. Self-Efficacy and Essay Revision Process in High School Writing Students). Broadly, theoretical studies describe specific mental and behavioral processes that underlie learning, and they are not studies in a computer science context, even if these studies have a learning theory bent (e.g. Constructivism in Computer Science Education).

Following Fincher and Petre's distinction between the direct and indirect use of theory, we further divided theoretical references into two categories:

- *Broad Theoretical*: The reference is to general literature in the learning sciences that has no subject specific reference point. This includes purely theoretical (e.g. Robbins (2004) Revision of achievement goal theory: Necessary and illuminating) as well as empirical studies, so long as they are general and not specific to an academic subject area (e.g. Green et al. (1989) Training in creative problem solving: Effects on ideation and problem finding and solving in an industrial research organization).

- *Analogy to Another Content Area*: The reference is used to bring some theoretical basis for understanding students' learning and behavior by analogy to a study from a specific content area outside computer science (e.g. Math Education). The purpose of the reference should be to posit an explanation or model for understanding student learning, but the content of the reference is a study that examines these processes or models in a specific non-CS content area (e.g. "Smith, Pasero, and McKenna [6] used data from Trends in International Mathematics and Science Study (TIMSS) and found that in fourth and eighth grade, boys had higher confidence than girls in science and were more likely to like science than girls.") This category is intended to refer to contexts in which authors make a general point about learning theory that is applied to their own CS context, but they do so using a study from another content area.

We also coded all articles for the presence of research questions. If we located within the article one or more questions that framed the purpose and scope of the study, whether or not they were explicitly labeled as research questions, the article was coded as containing research questions. Broader statements of research topic and aims that were not phrased as questions were not counted as research questions.

The articles were analyzed individually by three researchers. Each researcher coded a third of the articles from years 2013 through 2015, and 18% of the articles were used for inter-rater reliability pilot testing. For each pilot paper, two raters out of the three were chosen to code the paper independently for inter-rater analysis. The third rater coded the paper for inclusion in the data, after inter-rater reliability had been established. An iterative process of coding, revising the codebook, recoding, and discussion was used until the independent coders could consistently apply codes with an acceptable level of agreement on the pilot papers. The pilot papers were a subsample of the 2015 and 2014 articles. The 2012 articles were split between two of the researchers for coding, but were not made part of the inter-rater coding.

For subsequent coding rounds, codes were discussed and codes altered when one or more of the researchers believe they had miscoded their own work. When required codes (e.g. 'theory / no theory') were missing due to human error, these were flagged and recoded to ensure a complete data set. The most reliable results for inter-rater reliability were for the identifying the presence of learning theory code ($\kappa = 0.69$), whether or not a study was empirical code (κ =0.88), and the presence of research questions code(κ = 0.70). Due to the relatively low occurrences of the indicators of outcomes and methodologies, the respective κ values were biased towards larger confidence intervals, making them inappropriate for inclusion [22] [3].

7. RESULTS

The results of coding the articles are shown below. First, descriptive statistics are shown for all of the theoretical and methodological indicators used. Next, the results of comparison tests are shown for differences between CSE and ICER, differences between articles using theory and those not using theory, and differences between our results and the results of previous literature reviews. The test statistic refers to the independent samples z test for 2 proportions [12].

7.1 Use of outside Theory

The majority of the studies examined in both CSE and ICER made use of external theory. Descriptive statistics on the proportion of articles making use of outside theory are in Table 1. The majority of studies making use of theory used just direct references to theory while a smaller number used both direct references and references through another discipline, and almost no studies referenced external theory only through another discipline.

Table 1: Use of theory by publication

	Number of Articles	Yes Theory	No Theory
CSE	53	41	12
ICER	83	55	28

7.2 Methodological Quality

7.2.1 Randolph et al. Methodological Indicators

The results of the methodological analysis of the articles are in the tables below. The descriptive statistics for the methodological indicators, as defined in Randolph et al. [16] are shown first in tables 2 & 3, for CSE and ICER respectively, followed by the expanded set of indicators used in our expansion of the coding framework in tables 4 & 5, for CSE and ICER respectively. Note that Randolph et al. [16] used a specific definition of experimental to refer to any attempt to manipulate a variable and draw causal conclusions. Randolph et al.'s [16] definition encompasses all quantitative studies, regardless of whether they use a design that meets the strict definition of 'experimental', including studies from the single group posttest only. For this reason the first set of descriptive statistics provide a different value for experimental than the subsequent set, because for our coding scheme, we restricted use of the category empirical to its strict definition in which a variable is manipulated by random assignment of participants.

Table 2: Randolph et al. methodological indicators: CSE (empirical studies only)

	Num. Articles	Total	Prop.
Anecdotal	4	38	0.11
Attitudes Only	4	38	0.11
Sing. Group Posttest	8	38	0.21
Experimental	10	38	0.26
Qualitative	11	38	0.29

Table 3: Randolph et al. methodological indicators: ICER (empirical studies only)

	Num. Articles	Total	Prop.
Anecdotal	4	72	0.06
Attitudes Only	3	72	0.04
Sing. Group Posttest	20	72	0.28
Experimental	14	72	0.19
Qualitative	19	72	0.26

7.2.2 Additional Methodological Indicators

Descriptive statistics for the variables unique to our coding scheme are shown below, broken out by publication. Our indicators of type of research design and outcomes are shown in Table 4 and Table 5. Articles using quantitative research designs were coded for type of outcome, and a given study may have more than one outcome.

Table 4: CSE: Additional Methodological quality indicators and outcomes

	Num Articles	Total	Prop.
Research Questions	18	38	0.47
Mixed Methods	5	38	0.13
Single Group	15	38	0.39
Quasi-Experiment	10	38	0.39
Experimental	1	38	0.03
Attitudes	5	38	0.13
Psychological	9	38	0.24
Overall Grade	5	38	0.13
Data Mining	2	38	0.05
Other Outcome	17	38	0.45
Qualitative	11	38	0.29
Anecdotal	4	38	0.11

Table 5: ICER: Additional Methodological quality indicators and outcomes

	Num Articles	Total	Prop.
Research Questions	40.00	72.00	0.56
Mixed Methods	7.00	72.00	0.10
Single Group	30.00	72.00	0.42
Quasi-Experiment	14.00	72.00	0.42
Experimental	6.00	72.00	0.08
Attitudes	6.00	72.00	0.08
Psychological	11.00	72.00	0.15
Overall Grade	10.00	72.00	0.14
Data Mining	10.00	72.00	0.14
Other Outcome	23.00	72.00	0.32
Qualitative	19.00	72.00	0.26
Anecdotal	4.00	72.00	0.06

7.3 Comparisons between CSE and ICER

Results from our analysis were compared between CSE and ICER, to test the hypothesis examined in Randolph et al. [16], that articles from journals and conference proceedings differ in their rigor, using the five indicators as defined in that analysis. Also included are our indicators for the use of theory and the proportion of articles that describe an empirical study. The results of these comparisons are in table 6. These results show that CSE and ICER articles are no different in terms of methodological quality, confirming results reported in Randolph et al. [16]. However, the results of our analysis showed that ICER had a significantly higher proportion of articles describing empirical studies.

7.4 Comparisons between Theory and non Theory groups

After initial analysis, papers were grouped by whether or not they contained theory for the purpose of comparing these groups on other indicators. The intent was to examine whether rigorous use of theory correlated with rigorous methodological practices. The results of these comparisons are in table 7. These results show that articles making use of theory are significantly more likely to describe empirical studies, but all other comparisons show no relationship between theory use and methodological rigor.

7.5 Comparisons between current and previous results

Results of our article analysis were compared to those reported by previous literature reviews. However, a couple of caveats are necessary about how the data was equated between our study and the previous reviews. From Malmi et al. [10], we did not use the total proportion of articles they report as using a TMF, but rather, their reported proportion of articles using a TMF from education and psychology only. The results are in table 8.

For the comparisons with the Randolph et al. results, we calculated the proportions relative to the same reference groups as described in that study, which were different for each indicator. Contrast this to the descriptive results above, which were reported using the whole sample as a reference group. Furthermore, the comparisons to Randolph et al.'s results come with the limitation that our comparison was not identical to theirs. Whereas we focused on the most rigorous CS education journal and conference, Randolph et al. [16] used a cross section of CS education literature and did not provide results broken out by specific publications. For this reason, the reported differences can be attributed to the effects of differential rigor between publications, as well as changes in the field in the ten years separating our data from the previous study. These results are in table 9.

Lastly, we compared our results to those described in Al-Zubidy et al. [1], which categorized articles from the proceedings of SIGCSE with respect to whether or not they reported on an empirical study. Al-Zubidy et al.'s analysis used the same definition of experimental as used by Randolph et al. [16], so the comparison between our results and theirs are shown as well. These results are in table 10.

With the foregoing qualifications in mind, the results of our comparisons show some areas in which CS education research has changed in recent years. A significantly higher proportion of articles in our sample made use of outside theory than in the results presented by Malmi et al, suggesting that the field is increasingly reaching into other disciplines to frame and interpret studies with respect to previous research in learning theory. The comparison with Randolph et al.'s results shows that significantly fewer articles in our sample presented anecdotal or attitudes only results [16]. This could be the result of our limited breadth of analysis relative to Randolph et al., but results reported in Al-Zubidy, comparing their categorization SIGCSE papers to earlier reviews of SIGCSE papers, support the conclusion that CS education research is moving away from superficial results and more towards empirical results [1]. However, the results on the 'positive' methodological indicators (experimental and qualitative), as well as the 'negative' indicator of single group posttest only designs, show no change from Randolph et al.'s analysis. Furthermore, the prevalence of explicit research questions has increased significantly from the proportion reported in Randolph et al. [14]. One way to interpret these results is that while CS education is moving away from the superficial forms of research that it has been criticized for in the past, researchers have not yet moved towards toward significantly greater use of the more rigorous forms of non-superficial research. The comparison with the SIGCSE results give an interesting indication of the progress of the field as well. Al-Zubidy et al. noted that the proportion of SIGCSE papers reporting empirical results has increased dramatically over time [1]. Our results show that this progress has been substantial enough that the level of empiricism in SIGCSE papers has caught up to that of CSE and ICER. Nevertheless, the comparison of the proportion of experimental articles shows that there is still a gap in at least one indicator of methodological rigor between SIGCSE and the other publications.

8. DISCUSSION AND CONCLUSIONS

This literature review has attempted to give both a current picture of the rigor in CS education research as well as an idea of the trends across previous reviews and the current one. We also wanted to investigate the connections between use of theory and methodological rigor. Our review provides a description of the CS education literature showing that most research being published currently (in the venues we examined) is using theoretical constructs and frameworks drawn from general learning sciences research outside of CS education. This review also shows that a large majority of studies published are empirical. There are no benchmarks

Table 6: Methodological quality comparisons: CSE vs ICER

	test stat	p-value	prop CSE	prop ICER
Articles using Theory	1.92	0.17	0.77	0.66
Empirical articles	4.74	0.03	0.72	0.87
Experimental (Randolph sense) articles	0.69	0.41	0.26	0.19
Qualitative (Randolph sense) articles	0.08	0.77	0.29	0.26
Anecdotal (Randolph sense) articles	0.91	0.34	0.11	0.06
Attitudes only (Randolph sense) articles	1.69	0.19	0.11	0.04
Posttest only (Randolph sense) articles	0.59	0.44	0.21	0.28

Table 7: Methodological quality comparions: Theory vs no-Theory

	test stat	p-value	prop. Theory	prop. no-Theory
Empirical articles	4.34	0.04	0.85	0.70
Experimental (Randolph sense) articles	0.22	0.64	0.21	0.25
Qualitative (Randolph sense) articles	1.68	0.20	0.30	0.18
Anecdotal (Randolph sense) articles	0.00	0.98	0.07	0.07
Attitudes only (Randolph sense) articles	0.04	0.84	0.06	0.07
Posttest only (Randolph sense) articles	0.32	0.57	0.27	0.21

Table 8: Sample vs Malmi et al. (2014)

	Test Stat.	p-val	Sample	Malmi
CSE	27.07	0.00	0.77	0.33
ICER	15.11	0.00	0.66	0.36

Table 9: Sample vs Randolph (2007)

	Test Stat.	p-val	Sample	Randolph
anecdotal	35.23	0.00	0.07	0.38
attitudes	14.64	0.00	0.08	0.31
1 group post	2.48	0.12	0.37	0.49
experimental	2.21	0.14	0.74	0.65
qualitative	0.27	0.60	0.29	0.26
Research Qs	54.41	0.00	0.70	0.19
Empirical	10.14	0.00	0.81	0.66

Table 10: Sample vs Al-Zubidy et al. (2016) SIGCSE results

	Test Stat.	p-val	Sample	SIGCSE
Empirical	1.29	0.26	0.81	0.76
Experimental	9.12	0.00	0.55	0.38

for how much of CS education research should be empirical, but presumably, the greater the number of articles drawing from theory outside the field, the better.

Comparison of our results to the results of previous reviews gives cause for both optimism and further concern about the rigor of research in CS education. The increase in the use of theory from the learning sciences over the past few years is dramatic and demonstrates how quickly the field is developing. The reduction in the number of studies based on anecdotal reports or attitudes only data is a positive indicator of progress as well, as is the observed result that papers published in SIGCSE are also moving in this direction. However, the data on the other methodological indicators suggests that there is still work to be done to advance the rigor of research in CS education. While the superficial forms of CS education research have declined, the move of researchers towards more rigorous qualitative and quantitative designs has not been significant. These results are consistent with the notion that the field is in a gradual evolution. The first step in this evolution was to incorporate more theory from the learning sciences and move away from the less rigorous forms of research. These shifts in the research have served to partially answer the call made in Almstrum et al [2] for greater use of theory and rigorous methods. Nevertheless, the problem described by Robins [17], that most CS education researchers do not have the formal methodological training for rigorous behavioral science research, explains why the next step in the evolution has not occurred yet. Therefore, the next step in the development of CS education research should be a refinement of methodological approaches towards more rigor.

CS education research has long had concerns about rigor. This review has suggested that progress is being made. There are also important questions that go beyond the scope of this review. Questions raised by Malmi et al. [10] about the development of endemic theoretical constructs and frameworks in CS education remain unanswered. The question of how CS education compares in its rigor to other content area education fields, both now and historically, is also an interesting question for future research. Overall, however, CS education can be expected to continue to evolve methodologically as interest in the field continues to grow and researchers continue to gain more methodological training. Robins' [17] diagnosis of the field appears correct, and his advice for current doctoral students doing CS education work to pursue breadth in reading and methodological training seems to be an appropriate way forward to advance CS education as an increasingly well-developed field.

9. ACKNOWLEDGMENT

Part of this work was supported by the National Science Foundation under grant number 1502462. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

10. REFERENCES

[1] A. Al-zubidy, J. C. Carver, S. Heckman, and M. Sherriff. A (Updated) Review of Empiricism at the SIGCSE Technical Symposium. *Proceedings of the 47th ACM Technical Symposium on Computer Science Education (SIGCSE '16)*, pages 120–125, 2016.

[2] V. L. Almstrum, O. Hazzan, M. Petre, and M. Guzdial. Challenges to Computer Science Education Research. *Computer Science Education*, pages 191–192, 2005.

[3] T. Byrt, J. Bishop, and J. B. Carlin. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5):423–429, 1993.

[4] J. W. Creswell. *Research Design Qualitative, Quantitative, and Mixed Approaches*. Sage, Los Angeles, 3rd edition, 2009.

[5] J. W. Creswell. *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Pearson, Boston, 4th edition, 2012.

[6] S. Fincher and M. Petre. *Computer Science Education Research*. Taylor and Francis, 1st edition, 2004.

[7] M. Joy, J. Sinclair, S. Sun, J. Sitthiworachart, and J. Lopez-Gonzales. Categorising computer science education research. *Education and Information Technologies*, 14(2):105–126, 2009.

[8] Kirsti Klette. The Role of Theory in Educational Research. In *Norwegian Educational Research towards 2020 - UTDANNING2020*, pages 3–7, 2011.

[9] L. Malmi, J. Sheard, R. Bednarik, A. Korhonen, N. Myller, and A. Taherkhani. Characterizing Research in Computing Education : A preliminary analysis of the literature. *Sixth international workshop on Computing education research*, pages 3–11, 2010.

[10] L. Malmi, J. Sheard, Simon, R. Bednarik, J. Helminen, P. Kinnunen, A. Korhonen, N. Myller, J. Sorva, and A. Taherkhani. Theoretical underpinnings of computing education research. *Proceedings of the tenth annual conference on International computing education research - ICER '14*, pages 27–34, 2014.

[11] J. McMillian and S. Schumacher. *Research in Education: A Conceptual Introduction.* Longman, London, 2000.

[12] R. L. Ott and M. Longnecker. *Ott, R. L., & Longnecker, M. T. (2008). An introduction to statistical methods and data analysis.* Cengage Learning, Chicago, 6th edition, 2008.

[13] P. R. Pintrich and E. V. de Groot. Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, 82(1):33–40, 1990.

[14] J. Randolph, G. Julnes, S. Erkki, and S. Lehman. A methodological review of Computer Science Education research. *Journal of Information Technology Education*, 7:135–162, 2008.

[15] J. J. Randolph. *Computer Science Education Research at the Crossroads: A Methodological Review of Computer Science Education Research: 2000-2005*. PhD thesis, 2007.

[16] J. J. Randolph, G. Julnes, R. Bednarik, and E. Sutinen. A comparison of the methodological quality of articles in computer science education journals and conference proceedings. *Computer Science Education*, 17(4):263–274, 2007.

[17] A. Robins. The ongoing challenges of computer science education research. *Computer Science Education*, 25(2):115–119, 2015.

[18] J. Sheard, S. Simon, M. Hamilton, and J. Lönnberg. Analysis of research into the teaching and learning of programming. *Proceedings of the fifth international workshop on Computing education research workshop - ICER '09*, pages 93–104, 2009.

[19] P. Suppes. The Place of Theory in Educational Research. *Educational Researcher*, 3(6):3–10, 1974.

[20] A. Tellings. Eclecticism and Integration in Educational Theories: A Metatheoretical Analysis. *Educational Theory*, 51(3):277–292, 2001.

[21] D. W. Valentine. CS educational research: a meta-analysis of SIGCSE technical symposium proceedings. *ACM SIGCSE Bulletin*, 36(1):255–259, 2004.

[22] A. J. Viera and J. M. Garrett. Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37(5):360–363, 2005.