A (Updated) Review of Empiricism at the SIGCSE Technical Symposium

Sarah Heckman
NC State University
Raleigh, NC
sarah_heckman@ncsu.edu

Ahmed Al-Zubidy, Jeffrey C. Carver
University of Alabama
Tuscaloosa, AL
aalzubidy@crimson.ua.edu
carver@cs.ua.edu

Mark Sherriff
University of Virginia
Charlottesville, VA
sherriff@virginia.edu

ABSTRACT
The computer science education (CSEd) research community consists of a large group of passionate CS educators who often contribute to other disciplines of CS research. There has been a trend in other disciplines toward more rigorous and empirical evaluation of various hypotheses. Investigations of the then-current state of CSEd research showed a distinct lack of rigor in the top research publication venues, with most papers falling in the general category of experience reports. In this paper, we present our examination of the most recent proceedings of the SIGCSE Technical Symposium, providing a snapshot of the current state of empiricism at the largest CSEd venue. Our goal to categorize the current state of empiricism in the SIGCSE Technical Symposium and identify where the community might benefit from increased empiricism when conducting CSEd research. We found an increase in empirical validation of CSEd research to 72%; however, our findings suggest that current CSEd research minimizes replication precluding meta-analysis and theory building.

Categories and Subject Descriptors
K.3.2 [**Computers and Education**]: Computer and Information Science Education – *computer science education.*

General Terms
Measurement, Design, Experimentation, Human Factors.

Keywords
Systematic literature review, empirical computer science education, scholarship of teaching and learning.

1. INTRODUCTION
The SIGCSE Technical Symposium community is a large, welcoming, and vibrant group of passionate computer science educators. Most of us contribute to other communities of *discipline research* (e.g., software engineering) in computer science and beyond. However, many of the practices that we apply to demonstrate rigor in our discipline research are ignored or actively avoided when performing research in computer science education (CSEd) [2]. As evidence, Valentine [10] classified only 21% of CS1/CS2 papers published in the SIGCSE Technical Symposium between 1984 and 2003 as "experimental". Randolph, et al. [7] used Valentine's categorization and classified 40% of papers sampled from many CSEd venues between 2000 and 2005 as "experimental". However, the definition of "experimental" was broad and many papers reported results that preclude replication, meta-analysis, and theory building [4, 7, 11].

Researchers require evidence to determine the efficacy of teaching and learning interventions. Replications or comparisons of data across studies provide a basis for theory building [1, 4]. Fincher and Petre [4] describe two axes on which CSEd research can be classified: *evidence* – ranging from no evidence to empirical evidence and *argumentation* – ranging from low to high in argumentation or "theory." The authors argue that it would be desirable if the majority of CSEd research could be rated as high in empirical evidence to support theories, as we would expect in any other computer science discipline research [4]. However, they report that most CSEd research has evidence (possibly empirical) but little to no theory exploration [4]. Fincher and Petre [4] suggest that CSEd is "theory scarce" because most publications are not research and do not provide the evidence or replication required for meta-analysis and theory building. An increase in empiricism in CSEd research will move the field from "scholarly teaching" to the "scholarship of teaching and learning" (SoTL) [1] providing the foundation for meta-analysis and the generation of theories about teaching and learning in computer science [4].

Experience reports fall in the category of scholarly teaching rather than CSEd research or SoTL [1, 4]. Scholarly teaching is the application of current ideas or trends about teaching and learning usually due to reflection by an instructor on what did or did not work in the past [1]. SoTL moves scholarly teaching towards rigor, through a formalized plan and empirical evaluation, with the intention of peer review [1]. SoTL is the application of empiricism to CSEd that can build the theory for teaching and learning in computer science. We define empiricism as "validation based on observation of an intervention". An empirical validation reports results from observed evidence rather than argumentation, proof, or some other means [8]. Empirical validation is not solely experimentation or the scientific method, but incorporates the "method of science" [4]. The method of science considers both inductive and deductive paradigms for gathering evidence to answer a research question [4]. An increase in empiricism may help move CSEd research from a "soft" research area into a "hard" computer science sub-discipline [2].

In this paper, we present our examination of the most recent proceedings of the SIGCSE Technical Symposium, providing a snapshot of the current state of empiricism at the largest CSEd venue. Our goal is to categorize the current state of empiricism in the SIGCSE Technical Symposium [3] and identify where the community might benefit from increased awareness of empiricism when conducting CSEd research.

2. PREVIOUS CSED SURVEYS
Several researchers have explored the state of CSEd literature. Valentine [10] reviewed and classified the types of 444 papers about CS1/CS2 accepted to the SIGCSE Technical Symposium between 1984 and 2003. During the years surveyed, Valentine found that 21% of the reviewed papers contained some form of "experimental" evaluation. His experimental category was very broad and included papers where the "author made any attempt at assessing the 'treatment' with some scientific analysis" [10]. The

most common type of paper during the 20-year period was an experience report (what Valentine termed a "Marco Polo" paper). Valentine's data shows that the percentage of experience reports and experimental papers may vary from year to year. The overall trends show that CS1/CS2 experience reports were decreasing while experimental papers had a lull in the late 1980s and early 1990s and were starting to increase at the end of the survey period. However, these data may be skewed due to the selection of only CS1/CS2 papers and the timing of curriculum changes [10].

Valentine's [10] work has several limitations. First, while the period of evaluation was 20 years, he limited his paper selection to CS1/CS2 papers, which was between 25% on average of the proceedings between 1984 and 1993 and 30% on average of the proceedings between 1994 and 2003. A second limitation is that each paper was placed in a single classification when there may be multiple classifications that would work (e.g., there may have been experimental validation on an experience report). Valentine did all categorization himself; but other reviewers may have different interpretations of the categories. In our work, we address these limitations by allowing papers to fall into multiple categories and by allowing papers to be examined by two independent reviewers.

Randolph, et al. [7] reviewed a sample of literature published in multiple venues between 2000 and 2005 to understand the methodological properties of CSEd research during that time. They sampled 352 full papers from the June and December issues of the *SIGCSE Bulletin*, *Computer Science Education*, *Journal of Computer Science Education Online*, SIGCSE Technical Symposium, Innovation and Technology in Computer Science Education Conference, Koli Calling: Finnish/Baltic Sea Conference on Computer Science Education, Australasian Computing Education Conference, and the International Computer Science Education Research Workshop. Their findings suggest that much of the literature, 40% of studies with human subjects, limit their evaluation to anecdotal evidence. Of the 93 papers that reported an experimental or quasi-experimental design (less than one-third of the papers sampled), 54.8% used a post-test only design. The most common measure used to evaluate a paper's research goal was a questionnaire or survey, which consisted of 52.8% of the 123 papers that contained behavioral, quantitative, or empirical research. Randolph, et al.'s [8] work is limited in that several of the inter-rater reliability kappa statistics were low, but the key findings and those listed above were reported to have inter-rater reliabilities of good or fair.

Kinnunen, et al. [6] created a theoretical categorization of CSEd research with a didactic focus. They categorized 67 of the 72 papers published at ICER between 2005 and 2009 to identify commonly researched CSEd subjects. Kinnunen, et al. found that the most common categories of published work were related to students, including student's actions and understanding of learning outcomes, and to pedagogical activities used in the classroom. Of these papers, most were focused at the course level, with a few at the organization or society levels. There were no papers published in several categories suggesting that there are gaps in research coverage especially with teachers and teachers' interaction and engagement with students. The evaluation had at least two reviewers per paper, but no details or inter-rater reliability are reported. Kinnunen, et al. [6] were interested in identifying gaps in the literature for new studies. We are interested in characterizing the current state of the literature.

Prior work like Valentine [10] and Randolph, et al. [7] may have surveyed multiple years and multiple venues, but they did not consider the full proceedings for the SIGCSE Technical Symposium – only a subset. We are evaluating an entire proceeding as a snapshot of the current state of empirical evaluation in the SIGCSE Technical Symposium community. By considering the entire proceedings we can understand the community of the SIGCSE Technical Symposium, the flagship venue for CSEd.

3. METHODOLOGY

Our goal is to categorize the current state of empiricism in the SIGCSE Technical Symposium [3] and identify where the community might benefit from increased awareness of empiricism when conducting CSEd research. We are interested in the following research questions:

RQ1: What percentage of papers in the SIGCSE Technical Symposium have some form of empirical evaluation?

RQ2: Of the papers in the SIGCSE Technical Symposium with empirical evaluation, what are the characteristics of empirical evaluation in the papers?

3.1 Selection Criteria

Prior work [7. 10] has shown a small increase in empirical evaluation of papers in CSEd conferences in general and the SIGCSE Technical Symposium, in particular, between 1984 and 2005. Since we are interested in assessing the current level of empiricism in the SIGCSE Technical Symposium community, we evaluated all 110 accepted full papers in the Proceedings of the 45th ACM Technical Symposium on Computer Science Education [3]. We excluded panels, special sessions, posters, workshops, birds-of-a-feather, and keynote papers.

3.2 Data Extraction

For each full paper in the 2014 SIGCSE Technical Symposium proceeding [3], we determined if the paper contained any empirical evaluation. We defined empirical evaluation as evidence provided from observation, based on the software engineering literature's definition of evaluation (or validation) [8].

Our categorization metrics were chosen to assess the level of empiricism in CSEd research. The classifications of the relationship between the author(s) and evaluation subjects and data origins are a way of measuring the repeatability of the CSEd literature. Replicated evaluations provide a basis for meta-analysis and theory building.

If the paper contained an empirical evaluation, we then identified the following characteristics of the paper:

- Evaluation Type [7]: How were data collected?
  - Experiment: Data were collected through an experiment of some kind
  - Survey: Participants were surveyed about the intervention
  - Retrospective: Analyzing data that was previously collected for another purpose (i.e. not collected for the current study)
- Evaluation Subject [7]: What intervention was evaluated? Intervention types include: tool, pedagogical technique, curriculum, assignment, or other subject.
- Relationship between Author(s) and Evaluation Subject(s): Who developed the evaluation's subject?
  - Paper authors introduce a novel evaluation subject,
  - Paper authors use a novel evaluation subject from their own prior work,
  - Paper authors use a modified version of an existing evaluation subject (modified replication), and
  - Paper authors use an existing evaluation subject with no modifications (replication)

- Number of Students: How many students (if any) participated in the evaluation in the paper?
- Data Collected [7]: What was the actual data collected for the evaluation in the paper?
- Data Origins: Were the data collected for this evaluation or did the data already exist?
  - New: Data used in the evaluation was generated specifically for the evaluation
  - Historical: Data used in the evaluation was originally collected for another purpose
- Comparison [7]: Did the evaluation consider a comparison of the intervention to some other data set?
  - None: Contained no comparison.
  - New: Contained a comparison to something new or within the evaluation design.
  - Historical: Contained a comparison to existing data.
- Threats to Validity: Did the paper contain threats to validity about the evaluation?

Several of our categories are similar to categories used in related work. Table 1 maps the categories in Randolph, et al. [7] to categories that we used for evaluation type and subject. The idea of a control as a comparison is provided in Randolph, et al.'s research design category [7], but we are interested in determining if the control was part of the empirical evaluation or if existing data comprised the control.

3.3 Data Analysis

We reviewed each paper in the proceedings to determine if the paper contained an evaluation, and if so, identify the characteristics of the evaluation. Two reviewers, from different universities, conducted the review. One reviewer reviewed 60 papers and the other reviewed 50 papers. Each reviewer also reviewed a randomly selected ceiling of 25% of the papers from the other reviewer's set. There were 28 papers in common between the review sets. We used the reviewers' results on these papers to calculate inter-rater reliability. Like Randolph, et al., [7] we measured inter-rater reliability using a free-marginal kappa[1]. For each attribute described in Section 4, we report the kappa value resulting from the inter-rater reliability test to provide some indication of the level of agreement between the two reviewers. The strength of agreement ranges from a kappa value of less than 0.20 as poor agreement to kappa values of greater than 0.81 as very good agreement.

After the initial individual review, the reviewers met to discuss and resolve any discrepancies between their categorizations. There were some disagreements on some characteristics. Two papers required reading by a third reviewer (the third author) to resolve the categorization for one attribute. At the end, both reviewers agreed on the final categorization. After resolution, each reviewer revisited the other papers in their set to ensure that they did not need to make any adjustments before merging the results into a final set for counting purposes.

4. RESULTS

The results showed that 79/110 (71.8%) papers contained some form of empirical evaluation; over three times the percentage reported by Valentine [10] and not quite twice the percentage reported by Randolph, et al. [7]. From the initial evaluation of the papers, the kappa agreement rate was 0.7, which indicates a "good" level of agreement. The two reviewers were able solve all disagreements in a subsequent meeting. The following sections describe the results from characterizing the papers.

Table 1: Mapping of Categories to Randolph, et al. [7]

Source Categories	Our Categories
Evaluation Type	
• Experimental/Quasi-Experimental	• Experimental
• Qualitative	• Experimental
• Causal Comparative	• Experimental
• Correlational	• Experimental
• Survey Research	• Survey
Evaluation Subject	
• Course Org.	• Curriculum / Pedag. Tech.
• Tool	• Tool
• Teaching Programming Languages	• Pedag. Tech. / Assignment / Other
• Parallel Computing	• Other
• Curriculum	• Curriculum
• Visualization	• Other
• Simulation	• Other

4.1 Evaluation Type

To provide some insight into how researchers are evaluating their approaches, the first attribute is the type of evaluation used. Table 2 lists the evaluation types found in the papers along with the number of papers that used each one. Note that since many papers contain more than one type of evaluation, the total number of papers exceeds 79. The results show that Survey was the most common evaluation type. In fact, many of the papers that used an experiment or observation also conducted a survey. Most papers that Randolph, et al. [7] sampled with human participants were experimental (64.6%); the least were survey (7.6%). Since the SIGCSE Technical Symposium was the common venue, this may suggest that other venues have papers with more experimentation and fewer surveys leading to Randoloph, et al.'s differing results.

Table 2: Evaluation Type

Subject	# of Papers	Percent
Survey	46	58%
Experiment	26	33%
Others	7	9%
Observation	4	5%
Retrospective	3	4%
Pilot	3	4%
Quasi-Experiment	2	3%
Exploratory Study	1	1%
Not Reported	1	1%

Due to the possibility for studies to use multiple evaluation types, we calculated kappa values separately for each study type (i.e. did both reviewers agree on the presence of the study type). Of the study types listed above, only four were included in the common review set. The kappa values are Experiment (0.50), Survey (0.03), Retrospective Analysis (0.47) and Other (0.19). The kappa value for Experiment and Retrospective Analysis indicates a "moderate" level of agreement. The other kappa values indicate a "poor" level of agreement between reviewers. In analyzing the papers, we noticed cases where authors incorrectly named the type of

[1] We used MedCalc: http://www.medcalc.org/manual/kappa.php.

evaluation used (i.e. Pilot study and Quasi-Experiment). Furthermore, in some papers the lack of structure when describing the evaluation made it difficult to identify the evaluation type. This difficulty was likely the cause of the relatively low kappa values.

4.2 Evaluation Subject

The evaluation subject describes what intervention was evaluated: a pedagogical technique, a tool, a course or curriculum, an assignment, or some other subject. Table 3 shows the number of papers that evaluate each type of intervention. The total number of papers in Table 3 is greater than 79 because seven papers had more than one evaluation subject. The most common type of evaluation subject was pedagogical techniques; empirical validation methodologies are especially effective for evaluation of a pedagogical technique on students [1, 4]. Our top categories of pedagogical technique, course or curriculum, and tools correspond, roughly, to Randolph, et al.'s [7] top two categories of course organization (50%) and tools (19%).

Table 3: Evaluation Subjects

Subject	# of Papers	Percent
Pedagogical Technique	38	48%
Course or Curriculum	19	24%
Tool	17	22%
Other	9	11%
Assignment	3	4%

Because multiple Evaluation Subjects could appear in the same paper, similar to the Evaluation Type, we calculated a kappa value for each Evaluation Subject, as follows: Tool (0.37), Pedagogical Technique (0.14), Curriculum (0.03), and Others (-0.11). These kappa values indicate a "poor" level of agreement. We anticipate that the low kappa values were at least partially the result of the fact that many papers had more than one subject, which may have led one reviewer to identify one subject and miss the other. Additionally, the inconsistent paper structure in terms of how the evaluation process was reported made it difficult to correctly identify the Evaluation Subjects.

4.3 Relationship between Author(s) and Evaluation Subjects

If a goal for CSEd research is to build theory, then formalizing repeatable evaluation on common evaluation subjects (e.g., pedagogical techniques, course or curriculum, tool, assignment, etc.) will allow for meta-analysis and generalization. We categorized papers by the creator of the evaluation subject(s):

- Paper authors introduce a novel evaluation subject,
- Paper authors use a novel evaluation subject from their own prior work,
- Paper authors use a modified version of an existing evaluation subject (modified replication), and
- Paper authors use an existing evaluation subject with no modifications (replication)

Table 4 reports the results of the author of an evaluation subject. Each paper reporting an evaluation was only assigned one categorization. The kappa agreement between the two reviewers is 0.44, indicating a moderate agreement.

We found that the vast majority of paper authors introduced a new evaluation subject. Exceptionally few papers replicate prior work implying that there is very little comparison. Further, there is little replication of an author's own evaluation subject(s) in later work.

Table 4: Evaluation Author

Eval. Author	New/Mod./Rep.	# of Papers	Percent
Paper authors	New	62	78.4%
Existing	Modified	6	7.5%
Paper authors	Replicated	4	5.0%
Existing	Replicated	4	5.0%
Not Reported	Not Reported	3	3.7%

The lack of usage of previously established subject(s) and the lack of reuse of an author's own subject(s) makes it challenging to move the discipline of SoTL forward as it becomes increasingly difficult to compare the results of the various studies. A lack of replication precludes meta-analysis and theory building.

The lack of replication could be for a number of reasons:

- Faculty are moved from one course to another and do not have a good way to replicate their previous studies.
- Faculty are not able to recruit others to use their methodologies.
- Faculty tend to move to the "next great idea" when teaching their class and decide not to go back to their old methods.
- Faculty are unaware of other similar studies from which they could build their work upon.
- Published studies may not provide enough detail for replication by faculty.

4.4 Number of Students

One of the strengths of the SIGCSE community is the diversity of institutions members call home. Schools of all different sizes and backgrounds are represented by the over 1400 people that attend SIGCSE, which provides a large variety of institutions that can contribute each year. This diversity can be seen in the various papers that were reported in the 2014 SIGCSE proceedings.

Figure 1 shows the numbers of students used as subjects in the papers that reported an empirical evaluation. While the most common sample size was between 11 - 30 students, there is clearly a split between "small studies" and "large studies." We believe that this split occurs mainly based on the types of data that the researchers were able to use. For smaller studies ($n < 75$), the researchers are mainly looking at pedagogical techniques in smaller classes. The studies that reported more than 1000 students are retrospective studies using data collected for many years. The use of surveys for evaluation fell into both study sizes.

The kappa for this characteristic is 0.60, which is a "good" level of agreement. The disagreement between the reviewers on the number of students ranged from ±1 to ±15 because the number of students was not reported in a structured way. Some of the discrepancies are due to reporting the number of students in the studied population and then reporting a new number of participants due to some non-participation in the study. Multiple evaluations, each with their own set of students confounded agreement.

4.5 Data Collected

Different evaluations and evaluation subjects require different types of data. The papers included data such as: student assignments, quizzes, projects, lab assignments, exams, academic records, surveys, interviews, questionnaires, forums posts, and feedback posts. We grouped those various types of data into two groups: whether data were graded by an instructor (e.g. assignments, quizzes, etc.), or the data were feedback from the participants (e.g. questionnaires, forum posts, etc.). Table 5 reports the types of data collected and the number of papers that collected those data. The total number of papers is greater than 79 because

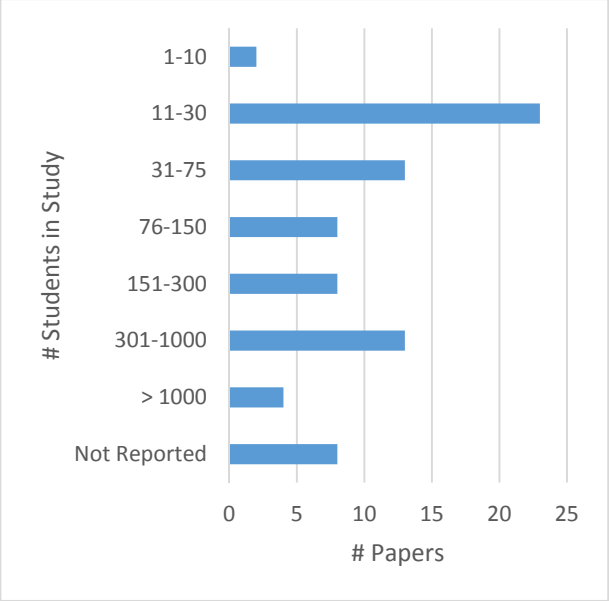

Figure 1: Histogram of Students in Evaluations

15 papers used multiple data sources. The kappa is 0.55, which is a "moderate" agreement. Randolph, et al. [7] collected similar, data; the most commonly collected items were questionnaires (53%) and grades (29%).

Table 5: Data Collected

Subject	# of Papers	Percent
Feedback data from participants	55	69.6%
Data graded by an instructor(s)	38	48.1%
Not Reported	1	1.2%

4.6 Data Origins

Data origins describes when the data was produced. Papers categorized as "new" report data that was generated specifically for the current evaluation. Papers categorized as "historical" report data that was collected for another purpose. Table 6 shows the data origins and the number of papers for each origin. No papers reported only historical data. Papers that collected a combination of new and historical data tended to have a comparison of the new data with the historical data (see Section 4.7). The kappa is 0.59, which is a "moderate" agreement.

Table 6: Data Origin

Subject	# of Papers	Percent
New	71	90%
Combination	8	10%
Historical	0	0%

4.7 Comparison

A comparison of an intervention with a control group or prior work strengthens the conclusions and provides a basis for theory building [4]. We classified the comparison into three categories:

- None: Evaluation contained no comparison between a treatment group and a control group.
- New: Evaluation contained a comparison between a treatment group and a control group where data for both groups was generated as part of the current study.
- Historical: Evaluation contained a comparison between a treatment group and a control group in which data for the control group was drawn from existing data that were not generated for the current study.

Table 7 reports the number of papers for each of these categories. Only 57% of the papers reported any type of comparison between the treatment and a control group. A lack of comparison to a control or prior work weakens the conclusions we can draw from those papers for theory building. Of the papers with comparison, approximately 2/3 compared the treatment to a control within the same study. The kappa is 0.44, which indicates a "moderate" agreement.

Table 7: Comparison

Subject	# of Papers	Percent
None	34	43.0%
New	29	36.7%
Historical	16	20.2%

4.8 Threats to Validity

A threats to validity or limitations section allows for readers to frame the results of the study. Almost 70% (55/79) of the papers did not report any threats to validity. All studies that involve human subjects have threats to validity due to choices made during the design process. One of the key aspects of study design, which makes it a difficult and time-consuming process when done correctly, is balancing various study designs with their inherent validity threats [4, 8]. It is possible to choose a study design that contains threats that completely invalidate the results of the study. But, more often the threats are less serious.

The biggest concern when papers do not report threats to validity is that readers are not able to properly interpret and apply the results [4, 8]. Without understanding the limitations of the study design, an educator may attempt to apply an intervention in an inappropriate setting, with potentially negative results [4, 8]. Authors should not think that reporting threats to validity weakens their paper, on the contrary it lends more confidence to the reader when he or she can understand the full context. The kappa value is 0.45, which is considered a "moderate" agreement.

4.9 Emerging Patterns

Overall, we discovered a general increase in the number of empirical studies in the 2014 proceedings versus the findings of others from the previous years. The increase may be attributed to our broad definition of empirical evaluation. An alternative explanation is that the review process for SIGCSE is looking more for papers with empirical evaluation, which in turn motivates researchers to include empirical evaluation in their studies.

However, the lack of reuse of evaluation subjects, or even the use of a modified version of a previously published evaluation subject, makes comparing results amongst contrasting pedagogical practices difficult at best. While the motivation appears to be in place to encourage empirical evaluation, researchers are effectively "reinventing the wheel" each time.

As noted in the discussion of many of the specific characteristics, inconsistent paper organization increases the difficulty of finding important information. Details about an empirical evaluation are not always reported in the same way, with the same information, or at the same level of detail. This lack of consistency makes it difficult to compare across papers and to get a full understanding of what really occurred during the study.

5. DISCUSSION AND RECOMMENDATIONS

Our evaluation of empiricism is not an argument to move the literature to require an evaluation, but a request for better evaluation when considering interventions. Replication of prior work is a key

practice in the advance of any scientific discipline and we see very little of it occurring in our sample precluding meta-analysis and theory building. There are many sources for information about conducting CSEd research that may serve as resources for the community [1, 2, 4, 5, 7, 10]. We advocate for a move from scholarly teaching to SoTL [1]. We propose the following concrete ideas for increasing empirical validation in CSEd research:

- Create an empirical validation or SoTL track at the SIGCSE Technical Symposium with an increase in paper length. Randolph, et al, [7] found that most CSEd research papers lacked a related work section, which may be the first item cut or minimized in a shorter paper. Longer papers would allow for more detail of methodology leading to additional replication. An empirical validation or SoTL track should emphasize replicative and theory-building work.
- Development of workshops co-located with the SIGCSE Technical Symposium and other CSEd venues to train educators in empirical validation techniques and provide feedback and mentorship on experiments.
- Building "laboratory packages" of classical CSEd empirical validation models [9], including qualitative models [5]. Many members of the CSEd community have large teaching loads, which would preclude the creation of a research methodology from scratch. Laboratory packages provide rapid adoption of stronger empirical methodologies and common metrics for replication and meta-analysis [9].
- Creation of datasets of student work for comparison and control.
- Move to consistent use of terminology and reporting of empirical results.

6. THREATS TO VALIDITY OF REVIEW

Our review of the 2014 SIGCSE Technical Symposium [3] full paper proceedings had several threats to validity. We mitigated one threat to internal validity by considering the full proceedings of the most recent SIGCSE Technical Symposium. However, that contributes to a weakness in our external validity: our conclusions are only about SIGCSE 2014 proceedings. With comparison to prior work, we can make inferences about the general trend of empiricism. Due to variations in the characteristics used to evaluate the literature, we cannot draw stronger conclusions.

A weakness in the construct validity is the inter-rater reliability scores for several of our characteristics. If recommendations for consistent terminology are taken for future work, then future surveys will have an increased construct validity. We considered the inclusion of a characteristic about specifically stating that the authors' Internal Review Board (IRB) approved the study, but decided that many papers may have left out the statement as assumed or due to space limitations. We assume that all evaluative work that used students as human participants received the necessary IRB approval.

7. CONCLUSIONS AND FUTURE WORK.

We have found that the reports of empirical CSEd research have increased when compared with prior surveys of the SIGCSE Technical Symposium. Seventy-two percent of the 2014 SIGCSE Technical Symposium proceedings [3] contained some form of empirical validation. However, many of the evaluations did not consider replication of methodologies or comparisons with other work. As CSEd research matures as a field, we must move toward meta-analysis of the literature for building of theories about CSEd. We plan to organize workshops at future CSEd venues for feedback and mentorship of empirical validation techniques. Additionally, we would like to contribute laboratory packages to help CSEd researchers with the empirical validation of their next teaching innovation.

8. ACKNOWLEDGEMENTS

We thank Kevin Lubick for serving as the second reviewer for the SIGCSE 2014 proceedings. We would also like to thank Lucas Layman of Fraunhofer Center for Experimental Software Engineering at the University of Maryland for his input on this work. Jeffrey Carver and Ahmed Al-Zubidy acknowledge partial support from NSF This material is based upon work supported by the National Science Foundation under Grant No. 1305395.

9. REFERENCES

[1] C. Bishop-Clark and B. Dietz-Uhler, *Engaging in the Scholarship of Teaching and Learning*, Stylus Publishing, Sterling, VA, 2012.

[2] T. Clear, "Valuing Computer Science Education Research?," Proceedings of the 6th Baltic Sea Conference on Computing Education Research, Koli Calling, Uppsala University, Uppsala, Sweden, 2006, pp. 8-18.

[3] J. D. Dougherty and K. Nagel, Conference Chairs, SIGCSE '14 The 45th ACM Technical Symposium on Computer Science Education, Atlanta, GA, USA, March 5-8, 2014.

[4] S. Fincher and M. Petre, eds., *Computer Science Education Research*, Taylor & Francis, The Netherlands, Lisse, 2004.

[5] S. Fincher, J. Tenenberg, A. Robins, "Research Design: Necessary Bricolage," Proceeding of the 7th International Workshop on Computing Education Research, Providence, RI, USA, August 8-9, 2011, pp. 27-32.

[6] P. Kinnunen, V. Meisalo, L. Malmi, "Have we missed something?: Identifying Missing Types of Research in Computing Education," Proceedings of the 6th International Workshop on Computing Education Research, Aarhus, Denmark, August 8-11, 2010, pp. 13-22.

[7] J. Randolph, G. Julnes, E. Sutinen, S. Lehman, "A Methodological Review of Computer Science Education Resarch," *Journal of Information Technology Education*, vol. 7, 2008, pp. 135-162

[8] M. Shaw, "Writing Good Software Engineering Papers," Proceedings of the 25th International Conference on Software Engineering, Portland, OR, USA, 2003, pp. 726-736.

[9] F. Shull, V. Basili, J. Carver, J. C. Maldonado, G. H. Travassos, M. Mendonça, and S. Fabbri, "Replicating Software Engineering Experiments: Addressing the Tacit Knowledge Problem," Proceedings of the 2002 International Symposium on Empirical Software Engineering, Nara, Japan, October 3-4, 2002, pp. 7-16.

[10] D. W. Valentine, "CS Education Research: A Meta-Analysis of SIGCSE Technical Symposium Proceedings," Proceedings of the 35th SIGCSE Technical Symposium on Computer Science Education, Norfolk, VA, USA, March 3-7, 2004, pp. 255-259.