On the Success of Empirical Studies in the International Conference on Software Engineering

Carmen Zannier
University of Calgary
Department of Computer Science
Calgary, AB, CAN
011 403 210 8153
zannierc@cpsc.ucalgary.ca

Grigori Melnik
University of Calgary
Department of Computer Science
Calgary, AB, CAN
011 403 210 9710
melnik@cpsc.ucalgary.ca

Frank Maurer
University of Calgary
Department of Computer Science
Calgary, AB, CAN
011 403 210 3531
maurer@cpsc.ucalgary.ca

ABSTRACT
Critiques of the quantity and quality of empirical evaluations in software engineering have existed for quite some time. However such critiques are typically not empirically evaluated. This paper fills this gap by empirically analyzing papers published by ICSE, the prime research conference on Software Engineering. We present quantitative and qualitative results of a quasi-random experiment of empirical evaluations over the lifetime of the conference. Our quantitative results show the quantity of empirical evaluation has increased over 29 ICSE proceedings but we still have room to improve the soundness of empirical evaluations in ICSE proceedings. Our qualitative results point to specific areas of improvement in empirical evaluations.

Categories and Subject Descriptors
D.2.0 [**Software Engineering**]: Experimentation

General Terms
Measurement, Experimentation

Keywords
Empirical evaluation.

1. INTRODUCTION
Empirical evaluation has received much attention in software engineering research in general, and in high profile software engineering conferences, such as the International Conference on Software Engineering (ICSE). This trend is based on critiques of the quantity of empirical evaluations in software engineering, as well as the soundness of empirical evaluations performed since the early days of software engineering research. This paper addresses both these issues. We present an empirical evaluation of peer-reviewed ICSE papers over the lifetime of ICSE.

Despite the 29 years of ICSE proceedings, empirically evaluated critiques of empirical evaluations are not seen. Keynote talks (e.g. [3]), panels (e.g. [24]) and workshops (e.g. [18]) support, instruct and debate empirical evaluations, but the results of such efforts are not yet known. This paper fills this gap by producing a first-time look at the evolution and growth of empirical evaluations within ICSE.

External to ICSE, numerous recommendations regarding how empirical evaluations should be performed can be found [6], [8], [13], and a few empirically evaluated critiques have also been performed [14][27]. These empirically evaluated critiques suggest that the percentage of papers containing empirical evaluations in software engineering journals and conferences is smaller than it should be. General opinion strongly suggests the same idea [8]. We sample ICSE proceedings, as the prime research conference in software engineering, to direct improvement efforts in software engineering empirical evaluation.

Given the context of our study we proceed as follows. Section 2 presents background wok and Section 3 presents the hypotheses of our study. Section 4 presents our experimental design via a discussion of the materials and procedure. Section 5 presents our results with some analysis. Section 6 focuses on analysis as it pertains to our hypotheses. Section 6 also discusses the highlights and lowlights of our results. Section 7 addresses the validity of our study and Section 9 concludes this paper.

2. BACKGROUND WORK
We present an overview of opinions on and recommendations for conducting empirical evaluations to motivate the selection of our evaluation criteria. First is the work of Basili [1][2][3][4][5]. Basili discusses the developing bodies of knowledge via an iterative model building, prediction, hypothesis testing, observation and analysis. [1][2][3][4][5] "Experimentation alone is of no value if there is no underlying framework where results can be interpreted." [2]. Concerning replication, Basili notes that the same experiment ran twice can yield different results because of the people involved. Thus, the importance of replicated studies is clear. "Experimental planning should have a horizon beyond a first experiment" [1]. "An analysis of several replicated projects can generate stronger statistical confidence in the conclusions" [2]. "We can ask if studies have been replicated under similar or differing conditions" [3]. "Too many studies tend to be isolated and are not replicated, either by the same researcher or by others" [4][5]. Given that these comments have been published in highly recognized journals for almost 20 years, it is remarkable that our results show a complete absence of replicated studies in our sample.

ICSE'06, May 20-28, 2006, Shanghai, China.

Second, we examine the work of Perry et al, [18], who begin by saying the problem with empirical studies is not in the details, but rather in the goals of those studies. Still, much of the paper specifies what a study should contain (i.e. details). Perry et al identify components of an empirical study as the research context, hypotheses, experimental design, threats to validity, data analysis and preparation and results and conclusions. They highlight the lack of hypotheses in empirical studies, and claim that we need to create better studies and draw more credible conclusions from them. Like Basili, Perry et al. refer to building a body of knowledge from empirical work. Given the strong emphasis on hypotheses in this work, it is interesting to note that further work [17], does not place as much weight on hypotheses.

A mini-tutorial by Shaw is regarding writing good research papers, which includes a component of "critiquing the lack of rigour in experimental software engineering" [21]. This work highlights the varying levels at which evaluation can occur and the ambiguity in defining these levels. The results for empirical evaluations were based on reading abstracts of the 2002 ICSE proceedings and in general, provide a more negative outlook on empirical studies than our results. Like Basili and Perry et al, Shaw emphasizes the importance of clarifying and explaining details of an empirical study.

A fourth prominent publication regarding empirical studies is the "Preliminary Guidelines for Empirical Research in Software Engineering" [13]. This paper provides explicit steps for the improvement of individual empirical studies to perform high quality meta-analysis. The experimental context, design, conduct, analysis, presentation of results and interpretation of results are all in need of improvement. For example, hypotheses should be made explicit as should a clear research question. Research related to the current study should be defined to evolve a body of knowledge in the given area. Biases should be reported for any work that is self-evaluated and the population to which the study applies should be made explicit [13].

Lastly, there are numerous evaluations of empirical studies in software engineering, that extend beyond ICSE proceedings. In [11] we see that the state of replication, theory revision and development of empirical software engineering methods is not as good as we might hope. The paper reviews content from the Journal of Empirical Software Engineering, over 6 years. Results indicate that a body of knowledge is difficult to define given a large spectrum of topics covered by empirical research [11]. Similar to [13], self-evaluations dominate empirical studies, and are considered to be "scientifically fraught". The paper summarizes concerns for empirical software engineering by stating, "experience shows that concentrating the investigation on just one of understanding or evaluation and without independent evaluation, replication and theory revision, it is unlikely that a useful theory will result." [11]". In [20] a similar study is conducted for the Journal of Empirical Software Engineering. The results indicated an ambiguous definition of the idea of software engineering topics. In [10] topics, research approaches, research methods, reference disciplines and levels of analysis are examined for six software engineering research journals. The paper concludes that the topics of software engineering experiments are diverse, the research approach and method are narrow as is the reference discipline and the level of analysis is technical. In [22] 103 articles are examined from 9 journals and three conference proceedings, over 9 years. The authors of [22] found a low number of controlled experiments, few industry-based studies and a lack of consistent terminology..

When we examine such comprehensive reviews of empirical software engineering from such prominent researchers we see much repetition in the need to develop a body of knowledge, conduct replicated studies, make explicit units of analysis, research questions and hypotheses, and clearly specify the population to which the study applies. However, given the numerous clear and repeated messages of [1-5][10-12][15-17] [23][26], which date back almost 20 years and provide results that date even further in history, we must ask ourselves, at what point will the message become clear?

We use a summary of the lessons learned from these prominent participants in software engineering empirical evaluation to establish the criteria by which we evaluate empirical studies. These criteria are defined in Section 4.2 and help to resolve the hypotheses we discuss below.

3. HYPOTHESES

We present two hypotheses each to be supported or refuted by our analysis. The first hypothesis concerns the volume of empirical work performed over the lifetime of ICSE. The second concerns the soundness of the empirical work performed. Our research hypotheses are labeled H_{11} and H_{12}. Their respective null hypotheses are labeled H_{01} and H_{02}.

H_{11}: The quantity of empirical evaluations performed has increased over 29 years of ICSE proceedings.

H_{01}: The quantity of empirical evaluations has not increased over 29 years of ICSE proceedings.

H_{12}: The soundness of empirical evaluations has improved over 29 years of ICSE proceedings.

H_{02}: The soundness of empirical evaluations has not improved over 29 years of ICSE proceedings.

Based on [4][11][13][18][26], we defined an empirical evaluation as being sound if it:

1. Makes clear four parameters: *Study Type*, *Sampling Type*, *Target and Used Population*, and *Evaluation Type*.
2. Implements legal (proper) use of a method of analysis depending on the scales of measurement.
3. Where appropriate, has well-defined hypotheses.

We define the following terms in Section 4.2: Study Type, Sampling Type, Target and Used Population, and Evaluation Type, Scale, Method of Analysis and Hypotheses.

In order to statistically evaluate our hypotheses, we compare early years of ICSE proceedings with later years. We divide our sample into two populations, the early years of ICSE up to and including 1990, and the later years of ICSE because we believe this is approximately when the push for empirical research gained momentum. It is also approximately the halfway point for ICSE, thus far. For our first research hypothesis, regarding the quantity of empirical evaluations in ICSE proceedings, we use a two sample test, [15]. The results from years 1975-1990 constitute the first sample population and the results from years 1991-2005

constitute the second sample population. We count the number of papers with an empirical evaluation component as a "success", represented as p_1 or p_2, for each population. Thus our first research and null hypotheses are:

$$H_{11} : p_2 > p_1 \qquad H_{01} : p_2 = p_1$$

Note the scenario $p_2 < p_1$ is included in H_{01} [15]. We selected a level of significance of .05 which is common for this type of studies [15].

For our second research hypothesis, we would like to use the Chi-squared goodness of fit test, but find that our sample size is too small. We provide descriptive results instead.

4. EXPERIMENTAL DESIGN

We outline the design of our experiment via a discussion of the materials and procedure used. For this study, the Study Type used is a quasi-random experiment, the Sampling Type is stratified random sampling, the Target Population is all peer-reviewed ICSE publications, the Used Population is peer-reviewed ICSE publications and the Evaluation Type is independent evaluation.

4.1 Materials

We randomly sampled ICSE publications across the 29 years of proceedings. The population from which the sample was drawn was peer-reviewed technical papers and experience reports over 6 pages in length. We did not include invited talks, panels, workshops or tutorials in the population. Our population size was 1227 papers. From this population we randomly drew 63 papers for our sample: just over 5% of the entire population. We used Java's random number generator to generate ID numbers corresponding to the papers. To ensure distribution across the 29 years, we grouped the proceedings into nine clusters of three year periods and randomly generated seven ID numbers within each cluster. We chose nine clusters of 3 years each because of basic mathematics. Cluster 1 groups ICSE proceedings 2005, 2004, and 2003, and we continue in this fashion to cluster nine, which groups ICSE proceedings 1978, 1976 and 1975 (there were no proceedings for 1977). This is a stratified random sample.

4.2 Procedure

We examined each paper to determine first if it contained an empirical evaluation component and second to determine if the empirical evaluation was sound. As per Section 2, we defined a sound evaluation as one that clarifies four parameters, states hypotheses where appropriate and performs legal analysis on the data collected. We define the four parameters now.

Study Type is the method of the empirical evaluation and can be any of a controlled experiment, a quasi experiment, a case study, an exploratory case study, an experience report, meta-analysis, an example application, a survey or a discussion. Table 1 defines each of these as found in the literature.

Sampling Type is the method by which the sample was chosen. The sampling can be any of simple random, stratified random, multi-stage random, non-random convenience, non-random self-selected, non-random investigator selected, non-random quota, non-random snowball, non-random purposeful or non-random critical case. Table 2 defines each of these as per literature.

Table 1: Study Type Parameters

Controlled Experiment
All of the following exist: Random assignment of treatments to subjects. Large sample size (>10 participants). Hypotheses formulated. Independent variable selected. Random sampling. [1]
Quasi Experiment
One or more of points in Controlled Experiment are missing. [3]
Case Study
All of the following exist: Research question stated. Propositions stated. Unit(s) of analysis stated. Logic linking the data to propositions stated. Criteria for interpreting the findings provided. Performed in a 'real world' situation [26]
Exploratory Case Study
One or more of points in Case Study are missing.[26]
Experience Report
All of the following exist: Retrospective. No propositions (generally). Does not necessarily answer how or why. Often includes lessons learned. [17]
Meta-Analysis
Study incorporates results from previous similar studies in the analysis. [9]
Example Application
Authors describing an application and provide an example to assist in the description, but the example is "used to validate" or "evaluate" as far as the authors suggest. [21]
Survey
Structured or unstructured questions given to participants. [16]
Discussion
Provided some qualitative, textual, opinion-oriented evaluation. E.g. compare and contrast, oral discussion of advantages and disadvantages.

The *Target and Used Population* is a 'yes' or 'no' answer to the question, does the target population match the used population? In other words, is the sample on which the evaluation was performed representative of the addressed population? As software engineering is targeting to help industrial software development, we assumed the target population was industry, unless otherwise stated by the author.

The *Evaluation Type* asks the question, who developed the subject under study and who evaluated it? Evaluation Type can be either of self-confirmatory or independent evaluation. In a self-confirmatory evaluation the authors play a large role in the production of the object of study (e.g. developed the tool or a method) and performed the evaluation. In an independent evaluation the authors evaluated a third party object.

In addition, we examined the papers for a 'yes' or 'no' answer to the question, are the hypotheses clearly stated? We also examined the papers for blatant illegal analysis on metrics. A prime example of this is calculating the mean on data from a Likert scale. Thirdly we determined if the evaluation contained any replication or was a replicated study. Lastly, we determined if the study reported primarily positive or negative results.

Our analysis using these measures was replicated internally. The first author evaluated all papers and the other authors validated this analysis by blindly evaluating 6 papers each, which were randomly assigned. This was an independent evaluation.

5. RESULTS

We provide results for the entire sample, then results for each cluster to show the improvement or decline of these results over the lifetime of ICSE. The Study Type, Sampling Type, Target and Used Population, and Evaluation Type, were evaluated from two perspectives: first from what the authors stated in their respective papers, and second from our perspective. Our perspective judged the four parameters based on the definitions provided in Section 3. When the paper did not provide enough information to state or infer the definition of each of the four parameters, the parameter was labeled Undefined (UD). If there was no evaluation component, the paper was listed as Not-Applicable (N/A). For all figures used in the results, the legend is as follows. AS means "Author Selected" and represents the values for the parameter from the author perspective. IS means "Investigator Selected" and represents the values for the given parameter from our perspective. No meta reviews were identified in any of the papers sampled.

5.1 Non-Applicable over Total

Of the 63 papers examined, 19 of them contained no evaluation component whatsoever. That is, 70% of the sample contained some form of evaluation.

5.2 Parameter Counts over Total

Of the 44 papers with some form of evaluation, we present the number of papers fulfilling each value for the parameters found in Tables 1 and 2 and for the values for Target and Used Population and Evaluation Type as described in Section 3.2. Figures 1 to 4 show the count for these values, under each parameter, across the lifetime of ICSE. We discuss each parameter separately. For the Author Selected perspective, the total count does not always equal 44 (e.g. Sampling Type). The remaining papers we marked as Undefined by the author. Sums from the Investigator perspective total 44.

Firstly, Experience Reports were the most common type of study performed, regardless of author or investigator perspective. Nine of the papers were author-claimed experience reports, and our analysis added 2 more to that group. Thus a quarter of ICSE papers containing evaluation components are experience reports. According to the authors, case studies were the next most frequently occurring evaluation. From our perspective, the term case study was frequently used improperly. Case studies lacked hypotheses, and/or a real world case. We found exploratory studies to be a better term for three studies claiming to be case studies. In total we found 6 exploratory studies.

Table 2: Sampling Types

Simple Random
Permits generalization from sample to the population it represents [16]
Stratified Random
The population is divided into a number of parts according to some characteristic. Increases confidence in making generalizations to particular subgroups.
Non-Random Convenience
Doing what is fast and convenient. On-the-spot decisions about sampling to take advantage of new opportunities during actual data collection [16].
Non-Random Self-Selected
Respondents decide they would like to participate in the experiment/case study/survey.
Non-Random Investigator Selected
The investigator selects the sample.
Non-Random Quota
Population is segmented into sub-groups, like stratified sampling. Then judgment is used to select the subjects or units from each segment based on a specified portion.
Non-Random Snowball
Identify cases of interest from sampling people who know people who know people who know what cases are information rich, that is, good examples for study.[16]
Non-Random Purposeful
Select information-rich cases strategically and purposeful; specific type and number of cases selected depends on study purpose and resources [16].
Non-Random Critical Case
Permits logical generalization and maximum application of information to other cases because if it's true of this one case, its likely to be true of all other cases [16].

While no controlled studies existed at all, we found 11 quasi studies, twice more than the authors clarified. The authors had a tendency to write merely that they had empirically validated or evaluated their work. They did not give their study a specific type. This initially seems like a small issue. However, much of the literature criticizes empirical studies for not communicating enough information about the published study so that it can be replicated [21]. Communicating the Study Type is clearly an important factor in replicating a study, thus we consider merely stating 'empirically validated' to be an insufficient description. We found significantly more agreement between the two perspectives for Evaluation Type, Population Type and Sampling Type, as shown in Figure 5. There was only one disagreement on Sampling Type and no disagreement on Target and Used Population, between Author Selected and Investigator Selected perspectives. Authors were more able to define their Sampling Type, Target and Used Population and Evaluation Type.

We found 5 studies were defined as a Discussion. This means they used primarily textual discussions to serve as the evaluation component of the paper. One study discussed the advantages and disadvantages of the presented concept. Another study compared the presented tool with 2 other similar tools. The other 3 studies provided a general discussion of the presented concept. We address Discussions in Section 6.2 in the context of varying types of evaluation methods. We found six papers that used an example as an evaluation. We address this in Section 6.2 also.

When we examined the type of sampling performed it became clear that a large number of the papers sampled (30 (68%) from the author perspective and 34 (77%) from the investigator perspective) were selected by the investigator of the study. This is well over two-thirds the papers containing an evaluation component. The remaining papers were evenly divided between convenience sampling and purposeful sampling, from both the author and the investigator perspectives. It is very interesting to observe that our random sample found no evaluations using random sampling of any kind. Investigator sampling is one of the weakest forms of sampling because it introduces an immediate bias into the study [16]. It is a type of convenience sampling but given the potentially high costs of empirical studies, researchers may be forced to work with what resources are available to them. Still, the imbalance in the sampling types is surprisingly large. Considering that empirical experimentation dictates that "only truly randomized tightly controlled prospective studies provide even an opportunity for cause-and-effect statements" [25], the majority of conclusions of the analyzed studies is only weakly supported.

The Target and Used population parameter highlights a positive point regarding the use of industry participants in studies. Half of the studies contained industry-related used populations in their study, both from the author perspective (23 out of 44) and the investigator perspective (24 out of 44). Given the time and resources often required to conduct industry-based empirical studies, we see this as a strength of empirical studies in ICSE proceedings. Such a number also shows that industry participants are also willing to support academic research, which is a positive step for empirical research. These results align with those presented in [11] regarding industrial data.

The opposite perspective of this is that half of the studies we examined did not have a target population matching their used population. If a non-industry population is used to conduct a study, the rules of statistics do not allow generalizing to an industry population. The situation intensifies when we recognize that in 10 of the studies the authors did not provide a clear enough definition of the population. From an investigator standpoint, we have 20 of 44 studies where the Target population does not match the Used population. Lastly, the Evaluation Type yields disappointing results. All but one evaluation was self-confirmatory. We address this further in Section 6.

5.3 Perspective Comparison

The results from Figures 1 to 4 show investigator and author perspectives of the sampled papers. It is hoped that as empirical software engineering matured, our ability to report on empirical studies aligned more with available guidelines for conducting empirical studies. We present our perspective as a summary of guidelines for empirical software engineering (and validate this in Section 7). Figure 5 shows the absolute count for each parameter over the lifetime of ICSE representing how many times our assessment agreed with the author's claims.

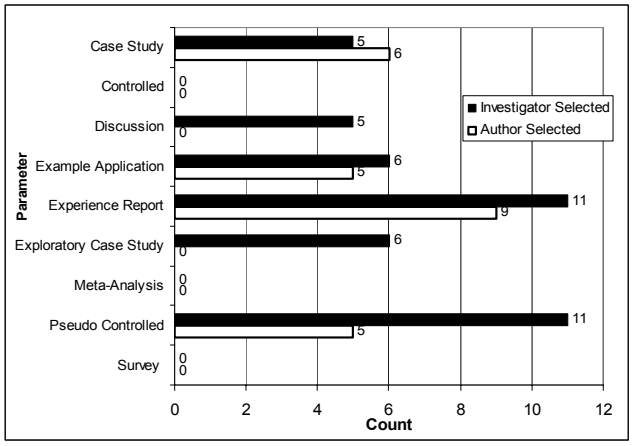

Figure 1: Study Type Totals of Sampled Papers with Evaluation Component

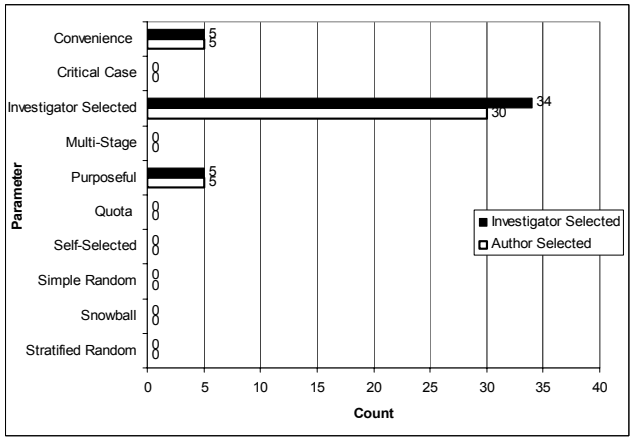

Figure 2: Sampling Type Totals of Sampled Papers with an Evaluation Component

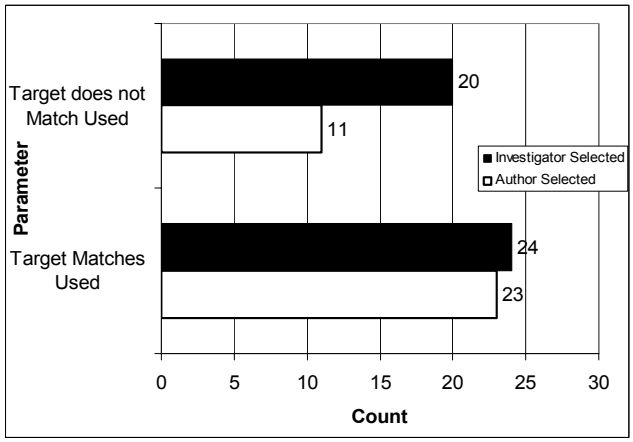

Figure 3: Population Type Totals of Samled Papers with Evaluation Component

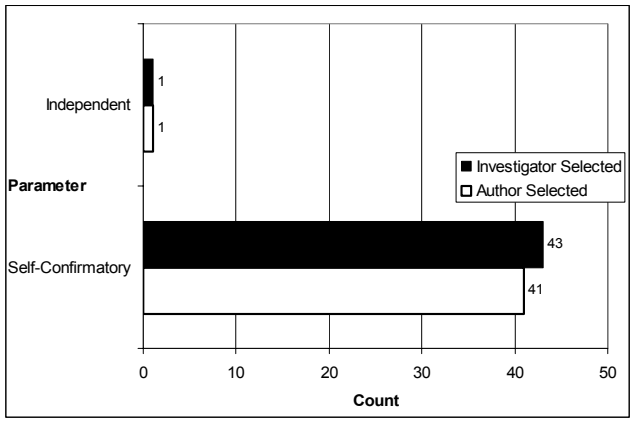

Figure 4: Evaluation Type Totals of Sampled Papers with Evaluation Component

From this figure two points are clear. Firstly, there is a significant lack of definition on the authors' part concerning the type of study they performed. This is shown by the AS = UD (Undefined) bar for Study Type in Figure 5. Almost half (18 of 44 studies) with an evaluation component lacked an explanation as to the type of evaluation they performed. Again, this makes replication of studies extremely difficult. It also leads us to question the extent to which the author understood exactly what was being evaluated and the validity of the results of the studies.

The second point represented in Figure 5 is the number of papers with an "incorrectly named" evaluation component. That is, we disagreed on the type of study actually presented by the authors 5 times. This is shown by the AS != IS bar for Study Type in Figure 5. For example, what one author called a case study, was in fact an experience report. This lack of agreement suggests we as a community do not agree on definitions of various empirical evaluations. There is a lack of consistent study type definitions in ICSE proceedings, despite published taxonomies such as [27].

5.4 First fifteen years vs. recent fifteen years

From the 63 papers examined and the 19 that contained no evaluation component, we now show the distribution of these papers across the 9 clusters. Figure 6 shows the number of papers per cluster with an evaluation component, of the seven papers sampled, per cluster. Over the lifetime of ICSE we see an increase in the number of papers published with an evaluation component. Statistically, we use a two sample z-test [15]. Our data is discrete but because we have a large sample size we are able to use a standard normal distribution [15].

Given a z* value of 1.96 and a level of significance of .05, using a standard normal distribution [15] we calculate a very low probability of a Type I error (.0256). Thus we reject the null hypothesis that $p_2 = p_1$. This strongly suggests that empirical evaluation is becoming more common. Overall, this can be seen as a maturing of the field.

Counts and Perspective per Cluster

Table 3We now examine the trends for the type of studies that have been published over the lifetime of ICSE. Table 3 shows the results from the author perspective and the investigator perspective, respectively. We know from Figure 1 that there is little agreement over the study types when taken as one large group. We therefore examine the results by cluster. Here, ideally we would see increasing similarity between the author and investigatory perspective as we move to the right of Table 3. This would mean that over the years the papers with an evaluation component were improving in consistency in defining the type of study. Without seeing increased similarity between the two graphs as we increase in years, we must acknowledge that the study type does not improve in its consistency of definition.

From both an author and an investigator perspective we see continued presence of experience reports across the years. From the investigator perspective we see a slight increase in the number of exploratory case studies, but this is not represented by the author's perspective. Our results also show a slight increase in the number of quasi controlled studies over the lifetime of ICSE, from the investigator perspective but this is also not recognized by the authors. This gives further credit to the idea that, as a community, we do not know exactly what types of studies we are publishing. In fact, when examining the author's perspective alone, we see little improvement in any area, as we increase in years. No Study Type recognizably improves in frequency and no real pattern exists across the years, except for the presence of experience reports. In general from Table 3 we see little consistency in what we are publishing. We have not improved in presenting our type of evaluation.

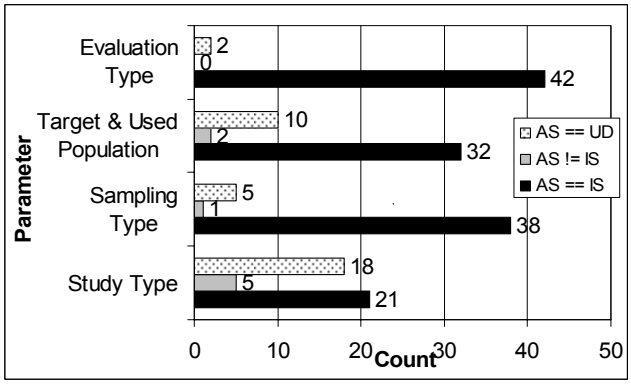

Figure 5: Total Agreements, Disagreements, Undefined

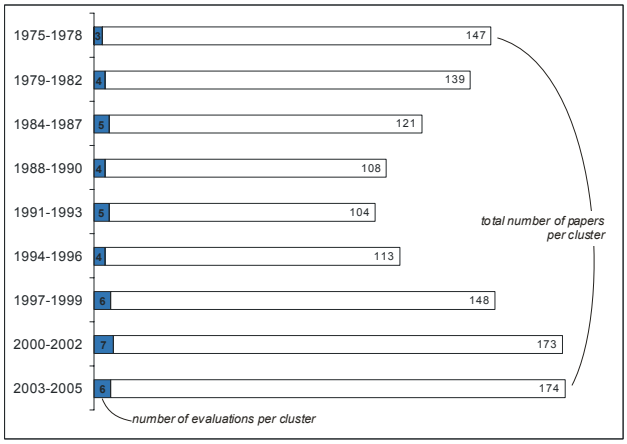

Figure 6: Number of Evaluations Sampled, per Cluster

Table 3: Study Type Trends, Author and Investigator Selected

Study Type Values	1975-1978		1979-1982		1984-1987		1988-1990		1991-1993		1994-1996		1997-1999		2000-2002		2003-2005	
	AS	IS	AS	IS	AS	IS	AS	IS	AS	IS	AS	IS	AS	IS	AS	IS	AS	IS
Case Study	**1**	0	0	**1**	0	0	**1**	**1**	0	0	**1**	**1**	0	0	**1**	**2**	**2**	0
Controlled	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Discussion	0	0	0	**1**	0	**1**	0	0	0	0	0	0	0	**2**	0	**1**	0	0
Example Application	0	0	0	0	**2**	**2**	**1**	**1**	**1**	**1**	0	0	**1**	**1**	0	**1**	0	0
Experience Report	**1**	**2**	**2**	**2**	**1**	**1**	**2**	**2**	**1**	**2**	0	0	0	0	**1**	**1**	**1**	**1**
Exploratory Case Study	0	**1**	0	0	0	0	0	0	0	0	0	**1**	0	0	0	**2**	0	**2**
Meta-Analysis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pseudo Controlled	0	0	0	0	**2**	**1**	0	0	0	**2**	**2**	**2**	**1**	**3**	0	0	0	**3**
Surveys/Interview	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Totals	2	3	2	4	5	5	4	4	2	5	3	4	2	6	2	7	3	6

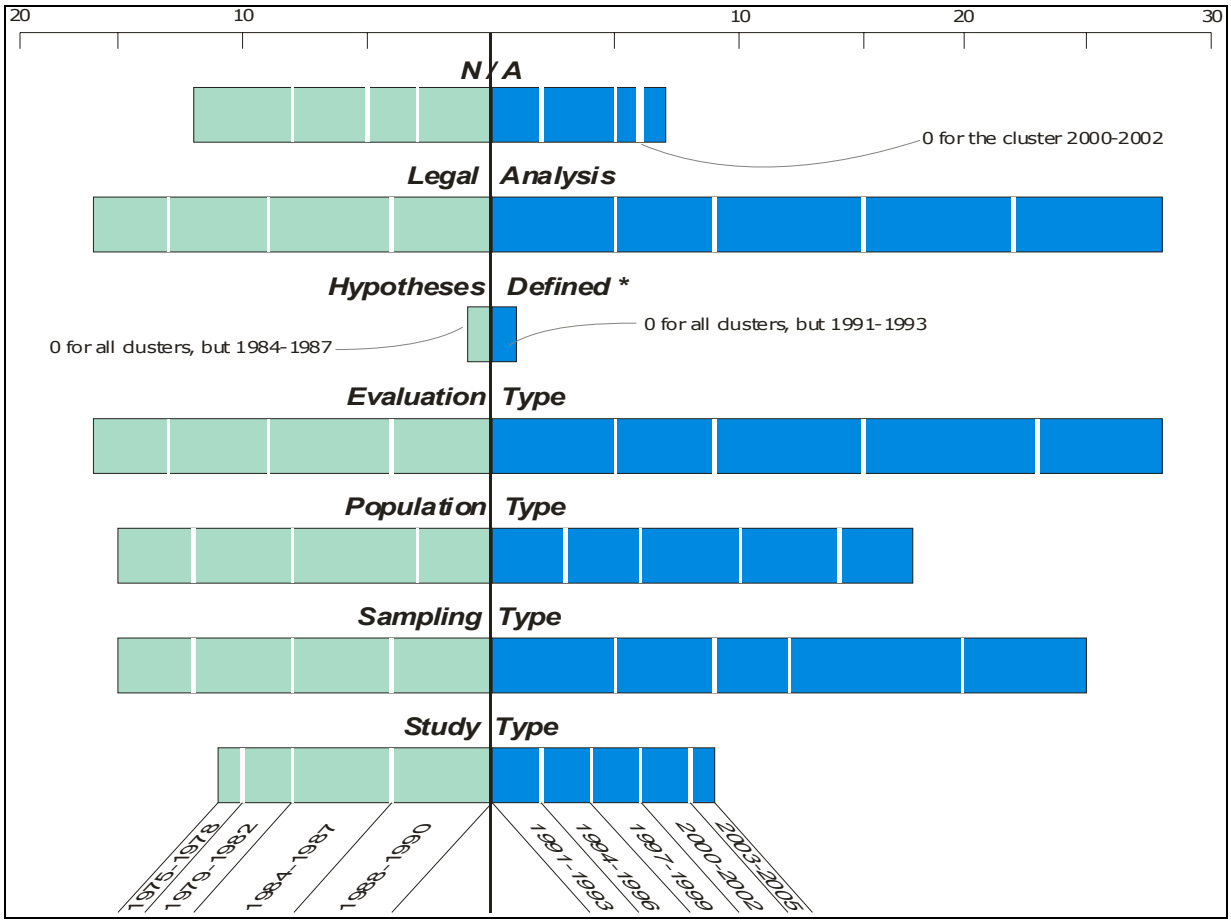

Figure 7: Number of Successes for Soundness Parameters. Left side for 1975-1990, Right side for 1991-2005

When we compare the two populations we see some improvement from the early years of ICSE to the later years. Figure 7 shows the number of successes for each parameter of soundness for the years 1975-1990 in comparison to the years 1991-2005. A success in soundness is denoted by an agreement between the author perspective and the investigator perspective. A Chi-squared goodness of fit test would be suitable here to statistically validate the extent to which the soundness is improving, but our sample size is too small for such a test [15]. From Figure 7 we see the number of successes improves slightly over the lifetime of ICSE, but we also note that the number of undefined parameters also increased. Figure 7 shows little improvement from the earlier years of ICSE to the later years, thus we have mixed feedback regarding soundness improvements over the lifetime of ICSE.

5.5 Hypotheses, Replication, Results, Analysis

We discuss hypotheses, replication, positive results and properly used analysis techniques in the context of the entire sample because there are no trends to report. Firstly, except for one study in our random sample, none of the examined studies contained hypotheses clearly stated. Secondly, none of the studies were replicated studies. Thirdly, except for one study, none of the examined studies reported primarily negative results. Lastly, all but one of the studies performed legal analysis on the provided data. The exception calculated the mean on ordinal data. We find these results quite astounding. Hypotheses are given much importance in textbooks [26][12], and in recommendations on empirical studies [18].

Given the severe absence of hypotheses formulation in our results, we propose three possible explanations. First, people conducting studies simply do not know that they should formulate hypotheses. A second possible explanation is that they feel their study is less formal than studies that require hypotheses. They potentially believe that stating a hypothesis requires the study to be more formal than it is. The last possible explanation is that those conducting the studies believe the formulation of informal hypotheses to be sufficient. We did not find any papers that formulated null and alternative hypotheses in a formal way.[1]

The absence of replicated studies contradicts the notion of accumulating a body of knowledge based on empirical evidence [6][18]. The absence of negative results (except once) questions the realism of the results we are producing (Is everything we do correct?) or the fairness of the conference review process (negative results do not make it through review processes).

6. ANALYSIS

In light of the quantitative results, we discuss our hypotheses, and lowlights and highlights of the results.

6.1 Evaluating the Hypotheses

We presented two hypotheses in Section 2. The hypothesis that the quantity of empirical evaluation performed has increased over 29 years of ICSE proceedings is supported by our descriptive and statistical results. Figure 6 shows the number of papers containing an evaluation component per cluster and we show a statistically significant improvement in the quantity of empirical evaluations over the lifetime of ICSE. Thus, we have rejected our first null hypothesis. This seems to be consistent with the observations in [18], [19] and [11] regarding growing awareness about empirical studies in software engineering.

Our second hypothesis, that the soundness of empirical evaluations performed has improved over 29 years of ICSE proceedings, cannot be statistically rejected at this time. From a descriptive standpoint, however, we see little improvement in the soundness of empirical evaluations, and thus qualitatively do not reject the second null hypothesis. Figure 5 shows results for the entire sample and shows undefined Study Type, Sampling Type, Target and Used Population and Evaluation type. This acknowledges the soundness of some studies can be improved. It also highlights some of systemic problems of the studies (such as lack of hypotheses) and of the state of the empirical practice in general (e.g. lack of replicated studies).

Table 3 focuses on study type over the lifetime of ICSE. We do not see increased agreement between author perspective and investigator perspective over the 29 years. This shows a lack of improvement in soundness in defining the study type. Figure 7 does not show a great deal of improvement over the lifetime of ICSE in terms of the number of successes, especially when one factors in the reduced number of N/A papers in the later years. This is a reasonable explanation for the higher number of successes in the second population and speaks to the quantity improving but not to the soundness improving. We emphasize though, that the number of successes for Sampling Type, Population Type, and Evaluation Type is better than the number of successes for Study Type. Given these results and existing background work that highlights inconsistent use of terminology in empirical studies, [10][20][22], it is difficult to specify study type definitively. This problem becomes even more apparent when we discuss the lowlights of our study results in Section 6.2. The outcome for our second hypothesis is not promising.

Our second requirement for sound empirical evaluation is that the evaluation performs legal analysis on scales of measurement. In this respect, our results are highly positive with only one occurrence of illegal analysis. The last requirement that appropriate hypotheses are well-defined, is not at all supported by our results. Given that the first and third requirements for soundness in empirical evaluations are not supported by our results over the lifetime of ICSE, we are forced to reject our second null hypothesis. Thus the soundness of empirical evaluation has not improved over 29 years of ICSE proceedings.

6.2 Lowlights

The rejection of the second hypothesis is addressed by four lowlights of empirical evaluation in ICSE, as per our sample. Firstly, our sample indicated a large misuse of the term case study. A case study answers a how or why question, is based on a research question and has data that links back to its research questions [26]. A case study is distinguished from an exploratory case study in that a case study has propositions (or hypotheses), an exploratory case study does not. A case study also occurs in a real world situation [26]. One concern, however, is that recommendations of how to perform empirical evaluations are giving mixed messages. Among ICSE published recommendations, hypotheses are given varying amounts of importance [1][18][17]. We include hypotheses here as a key role in a case study, as per [26], but also consider the extent to which the case studies contained an upfront research question, data linked back to a research question, and a real world situation. We recommend a consistent prescription and use of case studies especially versus exploratory case studies and experience reports.

Secondly, as shown in Figure 4, we see an extreme use of self-confirmatory studies. The pressure to publish new work, [21][5], is potentially an explanation for the gross self-confirmation bias in our results. This is highly ironic, given that a replicated study, seemingly not new, would be a new contribution to ICSE. Support for replicated studies is ubiquitous [11][4][5][19][13].

[1] We know that papers formulating hypotheses were published by ICSE but they seem to occur in so small numbers that they were not included in our random sample.

The third lowlight our results indicate is a misuse of the notion of an example. It has become clear that an example is still considered to be a type of validation [21]. We found phrases such as, "We validate our model with an example." We do not discourage the use of examples to clarify concepts. We discourage exploiting examples as evaluation.

Lastly, a Discussion also has a role in evaluation. Its current role is equal to that of a detailed case study or a quasi study. We recommend a distinction between empirical evaluation and some other term (e.g empirical discussion, empirical critique).

6.3 Highlights

Despite a seemingly bleak outlook on empirical evaluation in ICSE, our results highlight three points. Firstly, we have leaders in empirical evaluation who are examples of the soundness in evaluation that we can achieve. Researchers such as Basili, Briand, Jeffery, Kitchenham, Perry, Pfleeger, and Tichy participate in panels and keynote talks and continue to publish excellent work from which we can learn and establish benchmarks for empirical evaluation in ICSE. Secondly, our results do not show significant illegal use of analysis on scales of measurement, indicating that we as a community have already established some benchmarks and that the peer-review process of ICSE is a good one. Lastly, we cannot understate the increase in the number of empirical evaluations over the years. As a community we realize the importance of evaluation and are taking steps to accomplish it. Should this maturing trend continue, with increased knowledge of empirical evaluations from leaders of the community, there is little reason to believe the current state of empirical evaluations will not improve.

7. VALIDITY

We discuss the construct, internal and external validity of our study. We worked with a bias towards empirical studies, so if anything, our results are a positive reflection of the state of empirical studies in ICSE proceedings. We impacted our study by our interpretations of what the authors wrote in their paper. The potential for us to have misinterpreted something is mitigated by our internal replication.

7.1 Construct Validity

We acknowledge that our study works with the definition of a *sound* empirical study. We thus justify our definition of sound. In doing so we also justify the definitions of the four parameters we created. We compared publications on empirical studies from prominent researchers in empirical studies, to determine pertinent issues. The publications present recommendations for what an empirical study should contain. We noticed overlap among five publications and chose new names for each parameter for our review. Table 4 presents the recommendations from the literature and our version of the recommendations. Given that the authors of these publications are extremely well referenced (specifically in empirical evaluation), we believe this is a reasonable justification for our definition of soundness.

7.2 Internal Validity

To ensure we were consistent in defining a paper as a specific type we performed internal replication. We performed internal replication on 12 papers from the sample population. The results from the replicated study match those of the initial study showing a high likelihood that our results are internally valid.

Table 4: Construct Validity on Parameters

Our Terms	Yin [26]	Perry et al. [18]	Kitchenham et al. [13]	Basili [4]	Jeffrey [11]
Study Type	Research questions	Research Context; Experimental Design	"I3: Define the type of study."	Object of Study & Purpose	"wide ranges of empirical methods"
Sampling Type	Logic linking data, replication logic	Experimental Design; Threats to Validity	"D2: Define the process by which subjects and objects were selected."	Context of the Study	"students were the more common source of data."
Population Type	Criteria to interpret findings, replication logic	Experimental Design; Threats to Validity	"D1: Identify the population from which the subjects and objects are drawn."	"findings sometimes generalized to a population different from the sample."	"The balance … using industrial data [and] student data was very even."
Evaluation Type	Criteria to interpret findings	Experimental Design; Threats to Validity	"D8: make explicit any vested interests and report what you have done to minimize bias."	Point of View	"evaluations carried out by the inventor of the theory, which is scientifically fraught."
Hypotheses	Propositions	Hypotheses	"C2: …state [the hypothesis] clearly prior to performing the study"	Use GQM to evaluate appropriateness of a hypothesis.	"too little attention … to justification for the hypotheses"
Legal Analysis	Logic linking data	Data Analysis	"D4: Restrict yourself to simple study designs"	"measurements not always appropriate to goals of experiment"	"only implicit references to theory."

7.3 External Validity

The extent to which our study is generalizable to the population of all ICSE papers depends upon our sample size and external replication. Our sample size was over 5% of the entire population which was sufficient for statistical conclusions for our first hypothesis with a 95% confidence interval, and we believe, sufficient for the qualitative conclusions we make for our second hypothesis. The sample was drawn randomly. Replication beyond ICSE (e.g. to journal publications, where empirical evaluations are likely more prevalent) is future work.

8. CONCLUSIONS

We have provided empirically validated results of two research hypotheses regarding the current state of empirical evaluations in ICSE papers. We accept our first research hypothesis, that the quantity of empirical evaluations performed has increased over 29 years of ICSE proceedings. We reject our second research hypothesis and accept the null hypothesis, that the soundness of empirical evaluations has not improved over 29 years of ICSE proceedings. In terms of the first hypothesis, we report p_1=0.571 and p_2=0.8 showing the proportion of papers with an evaluation component is larger in more recent years than in earlier years of the proceedings of ICSE. We have a clear indication of an increase in the amount of empirical evaluations in ICSE. In terms of the second hypothesis, we see little consistent definition of the type of studies being performed and the lack of improvement in self-confirmatory studies strongly suggests that the soundness of empirical work has not improved. Additionally we see extremely little hypothesis specification. We are forced to reject our second research hypothesis and accept its null hypothesis. This paper presented a quasi-random experiment with quantitative and qualitative results to indicate the current weaknesses of empirical studies in ICSE. It was our intent to focus on issues in a specific community to focus improvement efforts in software engineering empirical evaluation.

9. REFERENCES

[1] Basili,V.R, et al "Experimentation in Software Engineering"; *IEEE Trans. Software Engineering* SE-12 7 July 1986.

[2] Basili; "The Experimental Paradigm in S/W Eng." *Proc. Int. Wkshp. Experiment. S/W Eng. Issues*; V706 1992.

[3] Basili V.R. "The Role of Experimentation in Software Engineering: Past, Current, and Future"; *Proc. 18th Int. Conf. S/W Engineering*; pp442-449, 1996.

[4] Basili V.R, et al; "Building Knowledge through Families of Experiments"; *IEEE Trans. Soft. Eng*. V 25 No 4, 1999.

[5] Basili V, et al; "Using Experiments to Build a Body of Knowledge" *Proc. 3rd Int. Andrei Ershov Memorial Conference on Perspectives of Sys. Informatics*; V1755; pp 26-282, 1999.

[6] Briand L; et al.; "Empirical Studies of Object-Oriented Artifacts, Methods and Processes"; *Empirical Soft. Eng*.; V.4 No 4, pp 287-404, 1999.

[7] Fenton N et al; Software Metrics: A Rigorous & Practical Approach 2nd Ed; PWS Pub. Company, 1997.

[8] Fenton N et al; "Science and Substance: A Challenge to Software Engineers"; *IEEE Soft*. V.11 No4, pp 86-95, 1994.

[9] Glass G.V; et al; Meta-analysis in Social Research; Beverly Hills, CA; Sage, 1981.

[10] Glass R.L et al; "Research in software engineering: an Analysis of the Literature" IST 44, pp491-506, 2002.

[11] Jeffery R, et al. "Has Twenty-five Years of Empirical Software Engineering Made a Difference?" *Proc. 9th Asia-Pacific Soft. Eng. Conf (APSEC 02),* 2002.

[12] Juristo N, et al; Basics of Software Engineering Experimentation; Kluwer Acad. Pub. Boston MA, 2001.

[13] Kitchenham, B et al., J; "Preliminary Guidelines for Empirical Research in Software Engineering"; *IEEE Trans. Software Eng,* V.28 No.8: 721-734, 2002.

[14] Lukowicz P, et al; "Experimental Evaluation in Computer Science: A Quantitative Study"; *J. of Sys. & Soft*, V.28 No.1 pp 9-18, 1995.

[15] Milton et al. Introduction to Statistics; McGraw-Hill, 1997.

[16] Patton M.Q; Qualitative Research & Evaluation Methods 3rd Ed.; Sage Publications, California, 2002.

[17] Perry D.E; et al; "Case Studies for Software Engineering"; *Proc. 26th Int. Conf. on S/W Engineering,* 2004.

[18] Perry D; et al; "Empirical Studies of Software Engineering: A Roadmap"; *Int. Conf. on S/W Engineering; Proc. of the Conf. on the Future of S/W Engineering*; pp 245-255, 2000.

[19] Pfleeger S.L; "Soup or Art? The Role of Evidential Force in Empirical Software Engineering"; *IEEE Software* Jan-Feb 2005 V 22 No. 1 pp. 66-73, 2005.

[20] Segal J et al. "The Type of Evidence Produced by Empirical Software Engineers"; *REBSE 05* St. Louis, Missouri, 2005.

[21] Shaw M; "Writing Good Software Engineering Research Papers" *Proc. 25th Int. Conf. on S/W Eng*.; pp 726-736, 2003.

[22] Sjoberg D.IK. et al; "A Survey of Controlled Experiments in S/W Engineering"; *IEEE Trans. Soft. Eng*.V31 #9, 2005.

[23] Tichy, W; "Should Computer Scientists Experiment More"; *IEEE Computer* V31, No.5 pp 32-40, May 1998.

[24] Walker R.J et al.; "Panel: Empirical Validation – What, Why, When and How"; *Proc. 25th Int. Conf. on S/W Engineering*; pp 721-722, 2003.

[25] Yancey, J. M; "Ten Rules of Reading Clinical Research Reports"; *American J. Orthodontics and Dentofacial Orthopedics*, V.109, No.5: pp 558-564, 1996.

[26] Yin R.K; Case Study Research: Design and Methods, 3/e Thousand Oaks, CA: Sage Publications, 2002.

[27] Zelkowitz M.V; et al.; "Experimental Validation of New Software Technology"; *S/W Eng. & Knowledge Eng; Empirical S/W Eng*.; pp 229-263, 2003.