Tangible assets to improve research quality: a meta analysis case study*

Alessander Osorio[1 [0000−0002−7576−0958]], Marina Dias[1 [0000−0002−8875−9438]], and Gerson Geraldo H. Cavalheiro[1 [0000−0002−4314−3429]]

Federal University of Pelotas, Pelotas RS, Brazil
{alessander.osorio, mldias, gerson.cavalheiro}@inf.ufpel.edu.br
http://www.ufpel.edu.br

Abstract. This paper presents a meta-analysis of the publications from all 18 previous editions of WSCAD in order to understand how performance results are validated and reported. This meta-analysis extract from these papers terms (keywords) belonging to three categories: statistics, metrics and tests. From all 426 papers analyzed, 93% referred at least one of the terms considered, indicating that there is a concern that results should be reported in order to the paper be considered relevant for this conference. Nevertheless, this analysis shows that only 3% of the papers applies reliable statistical tests to validate them. This paper depicts the meta-analysis achieved and proposes a direction to promote the adoption of a guideline to improve the results reporting in this conference and other with related subjects.

Keywords: Performance results analysis · Research evaluation · Statistical analysis · Research methodology.

1 Introduction

The proposition of a new technique or algorithm is usually followed by a performance analysis. It must take care to ensure that the performance study is performed so that the gain, if any, can be attested. It is usual, in experimental research, to dedicate a large amount of time to perform the experiments as well as a large amount of space on papers to present them, but performing a statistical study that validates the performance data comes second.

The *statistical study* considered in this paper is the one that results from the consistency and coherence analysis of the performance data obtained. Such study must precede the inference of behaviors, interpretations about the results collected and obtained conclusions. Therefore, the effective accomplishment of the statistical study allows associate reliability to the results presented in any scientific report.

In this work, a qualitative meta-analysis of a conference through an automated process of data mining was made. From the papers of the conference was

* This paper was realized with support of the National Program of Academic Cooperation from CAPES/Brasil.

extract the synthesis on the statistical and metric methods used. The objective of this work is present a landscape of how those papers statistically demonstrate their results, in order to provide indicators to qualify their submissions in the up coming years.

The case study is the Symposium of High Performance Computing Systems (WSCAD), a yearly Brazilian conference, starting in 2000, presented mostly in Portuguese. Where analyzed all 426 papers, wrote in Portuguese, from all firsts 18 editions (2000 to 2017). In addition, the present work also aims contribute to characterize, briefly, relevant statistical methods and techniques applicable for performance evaluation in high performance processing.

This work extends [17] bringing to the discussion ways of qualitatively improving scientific publications, specially in the domain of the statistical methods and techniques applicable for performance evaluation in high performance processing. Topics in statistics are explained to introduce basic concepts of data analysis and prove of measurements. Guidelines for the design, execution and reporting of research and its importance in the modern research scenario are also discussed.

The remaining of this paper is divided into 7 sections. The Section 2 characterize related works to the study presented in this paper. The quantitative analysis criteria are presented in Section 3 and the discussion about the methodology used is presented in Section 4. The results of the data obtained are discussed in Section 5. Section 6 are presented improvements of quality followed by Section 7 that have the conclusions of the paper.

2 Related Works

In this section we present some works developed in the context of the identification of methodologies of validation of results in scientific papers of the great area of Computation.

[18] evaluated 190 papers published on *Neural Networks Journal*, in the years 1993 and 1994 and showed that only 1/3 of the works had no quantitative comparison with previously known techniques. [22] analyzed 400 papers, published in ACM (Association for Computing Machinery), to determine whether computer scientists support their results with experimental evaluation. This study found that 40% of the papers did not have any type of evaluation. [23] reproduces the [22] research analyzing 147 papers published by ACM in the year 2005 concluding that 33% of papers are in the same situation. The work of [21] reinforces this evidence and cites the community's lack of experience in the correct analysis of the data in order to produce statistical evidence to prove the results as a cause.

Recently, the work of [1] study 183 papers from IPIN (International Conference on Indoor Positioning and Indoor Navigation) and concluded that although in many publications there was some concern in the evaluation of results, the quality of the description of the analysis methods was poor. Only 35% clearly report not only the methodology of the experiment itself, but what the results actually statistically represents.

Based on the results of the works above mentioned, the objective is to evaluate the publications of the Symposium of High Performance Computing Systems (WSCAD) focusing in how the statistical analysis of performance are described in their papers so that the gain can be attested, if there is.

3 Quantitative Analysis Criteria

Computer Science research, in most cases involves the development of a new application, algorithm or new computational system model [24]. Within this process, the research object is compared to similar techniques for effective performance evaluation of the proposed solution. This evaluation should be done by the quantitative analysis of the summarized results obtained by the use of synthetic data and statistical techniques of comparison of sets of measures.

Synthetic data, obtained by workloads, benchmarks, simulations and competitions, are classified into three categories. The first is used to evaluate the response time of a solution, the second to evaluate whether a solution can achieve the result (effectiveness) and the third to evaluate the quality of the response of the solution (efficiency) [24].

The performed experiments on a solution need to have effective statistical significance, according to the type of measurement performed and the appropriate statistical test to analyze this measurement. The types of measurement are categorical or nominal, ordinal, interval and reason [24]. Statistical tests are procedures used to test the null hypothesis, as the assumption is that there is no difference or relationship between the groups of data or tested events in the research object and that differences are due to random events, as well as the alternative hypothesis, the assumption is that there are statistically significant differences between measurements.

Calculating the probability of the null hypothesis being true or not, by the appropriate test according to the type of measurement performed, we found a number called p-value. When the level of significance represented by this value is lower than an indicator, with 0.05 (5%) being the most used value, the null hypothesis is rejected and the alternative hypothesis is accepted, that there is a difference and this one was not found randomly [24][3]. Although the hypothesis test is useful, when comparing values obtained in different experiments the hypothesis test is not enough. It is necessary to know how much these values effectively differ, using the so-called confidence interval. In [12], this confidence interval is set to at least 95%, which represents the largest and the smallest values assuming a p-value of 0.05 [24]. The confidence interval does not overlap, or invalidate, the standard deviation measure. The latter corresponds to the indication of how much the experiment data may vary from the mean and is used as a parameter in some tests.

The indicated, and therefore most commonly used, tests for comparing up to two sets of measurements and obtaining the p-value are: T-Test, Paired T-Test, U-test Mann-Whitney or Wilcoxon Rank-sum Test, Wilcoxon Signed-rank Test,

Chi-square, and Fisher's Exact Test. For multiple comparisons, with more than two sets of values, they are: ANOVA Test and Kruskal-Wallis [24].

Knowing that the statistical study should be applied to a collection of n performance samples collected, the remaining problem is how to define the value of n for a given experiment. The Central Limit Theorem (CLT) is the most important result in statistics, in which many commonly used statistical methods are based [14] to have validity. This theorem says that if a sufficiently large sample is drawn the behavior of the averages tends to be a normal distribution [14][2] or Gaussian [15]. The normal (or Gaussian) distribution is the statistical model that best represents the natural behavior of an experiment, where a random variable can assume any value within a defined range [15].

Sample here refers to the measurements of the research objects, number of replicates or iterations performed in the tests or experiments. Depending on the type of the research object, there are specific calculations for the sample size, however the CLT suggests that for most cases a sample size of 30 or greater is large enough for the normal approximation to be adequate [14].

There are applicable measurements criteria for each research object. According to [5] performance measurement can be classified as System or User-oriented measures. System-oriented measurements typically travel around the throughput and utilization. Transfer rate is defined as an average per time interval, be it tasks, processes or data. Usage is the measure of the time interval in which a particular computational resource is occupied. User-oriented measurements comprise response time and turnaround time.

Within this concept, it is possible to identify specialized metrics such as Reaction Time, Strech Factor, MIPS (Millions of Instructions Per Second), PPS (Packets Per Second), MFLOPS (Millions of Floating-Point Operations Per Second), BPS (Bits Per Second), TPS (Transactions Per Second), Nominal Capacity, Bandwidth, Usability Capacity, Efficiency, Idle Time, Reliability, Availability, Downtime, Uptime, MTTF, Cost/Performance Ratio [2][5] are inserted within these two generalizations.

The metrics terminology is quite extensive and diverse. It can also change according to the personal opinion of researchers or according to the country region. However, [2] and [5] illustrate both the definition and the use of each metric mentioned above.

4 Methodology

For data collect, all published works from 2000 to 2017 were saved locally, numbered and separated by folder according to the year of publication. Works in abstract format and those written in a foreign language were not part of the sample of the 426 analyzed publications. This is due to the fact that, given the nature of the abstract format, certain details of the data analysis of the studied object could be suppressed which could generate a bias. As well as works in foreign language would greatly increase the number of search terms if your diversity.

Following, through an automated software process (NVivo [1]), we searched for citations of terms, used in statistical analysis as well as metrics and tests used for gauging results, thus categorized in this article. These categories refer to good practices in data collection and analysis of results in scientific research, according to Tables 1, 2 and 3.

Description	Research Key
Sample (AM)	amostra
Standard Deviation (DP)	"desvio padrão"
Normal Distribution (DN)	"distribuição normal"
Frequency (FR)	"frequência" OR "frequência"
Gaussiana (GA)	gaussiana
Confidence Interval (IC)	"intervalo de confiança"
Mean (ME)	"média"
Num. Execution (NE)	"número de execuções"
Num. Iteration (NI)	"numero de iterações"
Test/Experiment (TE)	teste OR experimento OR simulação
Variance (VR)	variância

Table 1. Statistical terms selected for data collection

After organization, the data were summarized by year according to the category of the term. Tables 4, 5 and 6 show the results obtained in the search of statistical terms, metrics and tests respectively. Note that only terms that obtained results are summarized, those where there's no occurrences were suppressed. Still in Table 4, the results of the term frequency were suppressed due to the bias the results obtained. Frequency is quoted as the unit of measure of the processor clock. The results of the terms normal and Gaussian distribution were summarized together because they were the same object.

Table 7 summarizes the results from Tables 4, 5 and 6, as well as the distribution of citations by category of the term and year. In which is showed the total number of papers in the year, the number of papers containing at least one occurrence of the term. It also identifies the number of occurrences within papers separated by term category.

5 Discussion

The objective is not invalidate the results obtained in the experiments performed on papers that were analysed, but verify the care in publishing these results as well as the evolution of scientific research since they are almost 20 years of WSCAD. It is assumed that all possible errors and issues to avoid them have already been taken into account [2]. In this analysis, the proportion of papers with occurrence was used instead of the total of papers in the year and not the

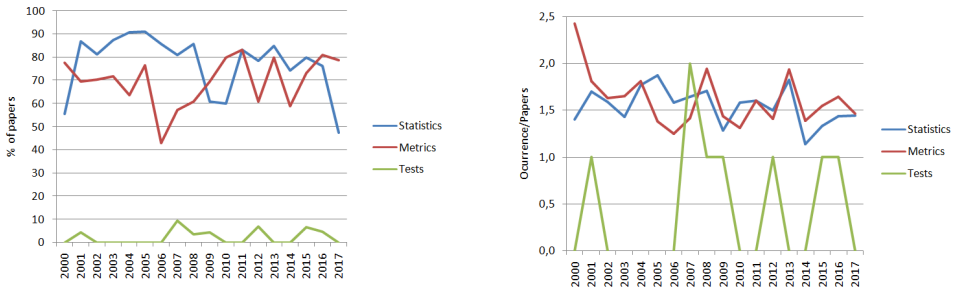
[1] http://www.software.com.br/p/qsr-nvivo

Description	Research Key
Bandwidth (BW)	”bandwidth OR ”largura de banda”
BPS (BP)	”bits por segundo” OR bps
Nominal Capacity (CN)	”capacidade nominal”
Usable Capacity (CU)	”capacidade utilizável”
Confidentiality (CO)	Confiabilidade
Performance (CP)	”performance”
Availability (DI)	disponibilidade
Downtime/Uptime (DU)	downtime or uptime
Efficiency / Accuracy (EA)	eficiência OR eficácia
Stretch Factor(FE)	”fator de estiramento”
Idle Time (TO)	”tempo ocioso”
MFLOPS (MF)	MFLOPS
MIPS (MI)	MIPS
MTTF (MT)	MTTF
PPS (PP)	PPS
Speed up (SU)	”speedup OR speed-up OR ”speed up”
Reaction time (TR)	”tempo de reação”
TPS (TP)	TPS

Table 2. Metrics selected for data collection

number of papers to allow the comparison between the years, since the number of papers per year is not the same having great variability Figure 1 (a). Of the 426 papers analyzed, 79% cited statistical terms, 67% metric and only 2% statistical tests.

The ratio was obtained by means of the data Table 7, dividing the number of papers with citations by the number of citations. The average citation of the three categories was 1.46 citations per article. Individually it is observed that for statistical terms (1,57) and metric (1,59) the values were above the general average in almost every year.



(a) Papers with occurrence by year (b) Occurrence by paper year

Fig. 1. Results Graphs

Description	Research Key
P-Value (PV)	"p-valor OR p-value OR "valor p"
ANOVA (AN)	anova
Chi-square (CH)	chi-quadrado OR qui-quadrado
Wilcoxon (TC)	"wilcoxon signed-rank"
Fisher (FI)	"teste exato de fisher" or "fisher"
Kruskal-Wallis (KR)	kruskal-wallis
T (TT)	"teste t" OR "teste-t" OR "teste de student" OR "Student"
U (TU)	"teste U" OR "mann-whitney" OR "wilcoxon rank-sum"

Table 3. Statistical tests selected for data collection

From the tests analysed, only T-Test or Student's test resulted in 9 occurrences, and two occurrences of the level of statistical significance (p-value), none of them were papers with the occurrence of Test T.

Due to the low occurrence of statistical tests in the results, around 2%, it was decided to investigate this data more closely. For this, a second sample from the total 426 papers was calculated, disregarding the 9 studies where the occurrences of these terms were already found, with a 95% confidence interval and a sampling error of 5%, reaching a total of 30 randomly selected papers.

The second sample was sent to two reviewers, who read and analyzed the methodology and analysis of the results of these papers according to some criteria. Experimentation, whether carried out or not; Sample method referring to the method used to find the sample size, by the use of benchmarks, estimation or calculation; Size of the sample represented by the number of experiments performed, repetitions, instructions, jobs and the like; If the size of the sample is evident and if its size is adequate; Use of metrics; If there was a comparison with another technique and this technique was adequately demonstrated in some way.

From the papers analyzed in the second sample, none showed equality or deficiency in their results, the worst case only points to "evidence of performance improvement." From this sample, 24 of them performed experimentation, 6 are theoretical models or descriptive memorials of implementations of computational systems or solutions, without numerical analysis of results. Of those who carried out experimentation 10 they used benchmarks, 13 estimated the sample size and 1 did not quote numbers.

Sample sizes were considered adequate for all those who used benchmarks since they are large collections of records and it is implied that for each input record there is an output measurement and recording, also because they are a method widely accepted in the scientific community [24]. From those studies where the sample was estimated (13) in 11 of them the sample size was considered adequate, given the sample size and in 2 it was considered inadequate because they were below the minimum stipulated by CLT [14]. It is necessary to be aware, since a benchmark can represent the input load to which an experiment is submitted and not the amount of data collected, looking at this aspect only

Year	n	AM	DP	DN	IC	ME	NE	NI	TE	VR
2000	7	-	-	-	-	2	-	1	4	-
2001	34	4	1	-	-	10	1	1	16	1
2002	35	1	2	-	-	9	1	2	20	-
2003	40	2	3	-	-	11	1	2	21	-
2004	53	-	4	-	-	19	1	4	24	1
2005	58	3	6	-	1	17	-	2	28	1
2006	38	-	2	1	-	7	1	3	24	-
2007	28	-	1	-	-	7	1	1	17	1
2008	41	1	3	1	1	12	-	1	22	-
2009	18	-	-	-	-	4	-	-	14	-
2010	19	2	-	1	-	5	-	-	11	-
2011	8	-	-	-	-	3	-	-	5	-
2012	33	-	1	1	1	7	1	-	21	1
2013	31	2	2	2	-	11	-	-	14	-
2014	33	-	-	-	-	3	-	2	27	1
2015	16	-	1	1	1	1	-	1	11	-
2016	23	1	-	1	1	2	-	1	16	1
2017	13	2	1	-	-	2	-	-	8	-

Table 4. Citation number of Statistical Terms by year

8 papers had adequate samples with statistical analysis power specified in the text.

In all papers where there were experiments (24) were metrics, from this total 15 compared their results to other techniques and 9 did not. Confirming the results for using metrics from the first sample at 70%.

The second sample show only in 3 papers were the results are based on scientific criteria (statistical methods and mathematical models), but only 1 article minimally used statistical methods (p-value and DP) to confirm their results. The values for the first and second sample evaluation criteria are similarly 2% and 3% respectively.

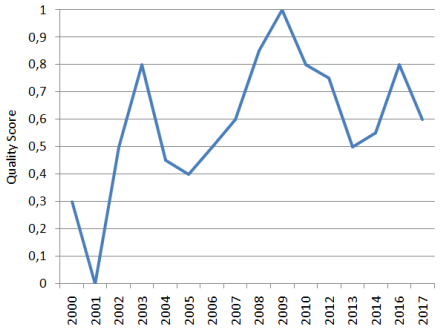


Fig. 2. Quality Score per Year

Year	n	Disp.	BW	BP	CP	DU	EA	ID	MF	MI	PP	CO	SU
2000	17	5	1	-	-	-	6	1	-	1	-	1	2
2001	29	7	3	-	-	-	9	-	-	1	-	2	7
2002	31	5	1	-	-	-	12	-	-	3	-	4	6
2003	38	11	2	-	-	-	8	-	3	2	-	5	7
2004	38	10	-	-	-	-	14	-	-	1	-	6	7
2005	36	10	1	-	-	-	9	1	-	1	-	4	10
2006	15	3	2	-	-	-	4	-	-	1	-	-	5
2007	17	3	1	-	-	-	3	1	-	2	-	-	7
2008	33	6	5	-	-	-	7	-	1	3	-	3	8
2009	23	6	1	1	-	-	4	1	1	1	-	1	7
2010	21	5	-	-	-	1	2	1	-	5	-	2	5
2011	8	3	-	-	-	-	1	-	-	1	-	-	3
2012	24	9	2	1	-	-	4	-	-	4	-	-	4
2013	31	5	-	1	-	3	5	1	-	3	-	2	11
2014	32	6	-	-	-	-	5	1	-	3	-	1	16
2015	17	1	-	-	-	-	3	1	1	2	1	-	8
2016	28	6	3	-	1	-	1	4	1	1	-	1	10
2017	22	5	1	-	-	1	-	2	-	1	-	1	11

Table 5. Citation number of Metrics Terms by year

Year	n	PV	TT
2001	1	1	-
2007	4	1	3
2008	1	-	1
2009	1	-	1
2012	2	-	2
2015	1	-	1
2016	1	-	1

Table 6. Citation number of Tests by year

To calculate the evolution of the WSCAD in terms of quality of publication, each positive result of the categories analyzed in the second sample was assigned weight 1, and zero for negative or nonexistent results. The sums of each year and averages were calculated in a score which comprises values between 0 and 1. Figure 2 shows the evolutionary line of the score of quality of the publications, evidencing the maturation of the event over time. Although there are alternations in the line, the trend of the curve is upward with most years above average.

6 Improving the quality

As presented in Section 5, statistical tests are not so cited as terms and metrics. The assumption, is that this may happen for two reasons. First the statistical test was not effectively applied to obtain the results, and second, due the fact that it not described in the paper methodology or discussion sections. To improve

Ano	n -	Statistic		Metric		Test	
		Art.	Cit.	Art.	Cit.	Art.	Cit.
2000	9	5	7	7	17	-	-
2001	23	20	34	16	29	1	1
2002	27	22	35	19	31	-	-
2003	32	28	40	23	38	-	-
2004	33	30	53	21	38	-	-
2005	34	31	58	26	36	-	-
2006	28	24	38	12	15	-	-
2007	21	17	28	12	17	2	4
2008	28	24	41	17	33	1	1
2009	23	14	18	16	23	1	1
2010	20	12	19	16	21	-	-
2011	6	5	8	5	8	-	-
2012	28	22	33	17	24	2	2
2013	20	17	31	16	31	-	-
2014	39	29	33	23	32	-	-
2015	15	12	16	11	17	1	1
2016	21	16	23	17	28	1	1
2017	19	9	13	15	22	-	-

Table 7. Distribution of citations by type of term and year

the results quality is necessary avoid both issues, applying the adequate test to calculate the results, and reporting how it was done. Next in this section, is discussed some statistical subjects and report strategies to try mitigate these issues.

6.1 Subjects in Statistics

The absence of the statistical analysis in the results verification could be an indicator that it was not performed. Second [21] one of the reasons for this, is the lack of knowledge of the authors for the execution of this stage of the research. In an attempt to mitigate this situation, the basic concepts in statistics are described below.

Normal Distribution or Gaussian (DN) - [7] describes the normal distribution, how the most commonly used distribution in data analysis. [12] enrols DN as a central role of data analysis. As already described in Section 2, DN is the model that best represents the natural behaviour of an experiment. Graphically the DN is represented by a symmetric bell curve, in which the bigger part of the sample values is around the center of the curve (mean) with some variability. It's important know about DN, because many of the statistical procedures like correlation, regression, t-tests, and analysis of variance needs that the data follows a DN [7].

Central Limit Theorem (CLT) - says that if a sample is large enough the behavior of the values tends to be a normal distribution [14][7][2]. The CLT suggests that samples with 30 or more observations are sufficient to have a

normal distribution of the data. The CLT allows that a non-normal distribution population be sampled to a DN by a representative sample of this population, in this way statistical procedures that needs a DN of data can be executed.

Sample (AM) - is a part of a population. Population is a complete universe of all individuals that will be studied [7]. The statistics objective is make inferences of a population by using samples of this one [12]. For this reason is important distinguish sample from population. In most of cases it's impossible, or expensive, study the whole population. So a representative sample is needed, keeping in mind the CLT statements not only for sampling, but with the number of executions (NE) or iterations (NI).

Hypothesis Test (HT) - most of experiments are trying to proof something. For example: "there are improvements between technique A in relation to B?". This sentence is the hypothesis statement. The *null hypothesis* (*H0*) is equality hypothesis, and the *alternative hypothesis*(*H1*) are the opposite. In this case the *H0* is that there are no improvements from technique A to B, and the *H1* is that A have improvements over B. The HT works rejecting or accepting one or other hypothesis by calculating the *p-value* using an appropriate statistical test and a significance level (1% or 5% generally used). If the *p-value* is smaller than the significance level used in to calculate, *H0* is rejected and *H1* accepted (there are significant improvements), otherwise, *H0* is accepted and *H1* rejected (there are no significant improvements).

Confidence interval (IC) - adding the concepts for confidence interval listed on Section 2, the IC is commonly used to indicate reliability from research results. Note that the more critics need to be the search results the greater the confidence interval. The pharmaceutical industry, for example, adopts a 99% IC in drug tests (p=0.01). In some cases not of non-critical result studies, more relaxed values of IC are acceptable, however the scientific community adopts the value of 95% (p = 0.05) as standard.

Mean (ME) - is one measure of summarization of the sample. Is obtained by the quotient of summarizing the data by sample size. It's important distinguish mean from other two summarization measures: median and mode. Median is the central point of a sample data, that separates the sample in two halves. Mode is the value of a sample with major occurrence. These measures in a DN tend to be the same.

Standard Deviation (DP) - is one dispersion measure and your values have the same unit of the data. Your results represents how far (big DP values) or close (smaller DP values) to the ME of the data. Represents the variability of the sample and is used for many other calculations in statistics, for example IC.

Variance (VR) - is another dispersion measure, similar to the standard deviation. His unit is expressed in the square of the unit of the sample measurement. For this reason it is not commonly used, being its square root, the standard deviation, most used to represent the variability of the data. Variance is more used to compare when performing variance analysis of more than one set of values (ANOVA).

P-Value (PV) - is the probability that the test statistic will take on a value that is at least as extreme as the observed value of the statistic when the null hypothesis H0 is true [12] in other words represents the lowest level of significance which null hypothesis is rejected. PV is the result of most of statistical tests. His value helps to improve the power of research results.

With the advance of many statistical software packages in the market nowadays, knowing how exactly a test is calculated is not (sometimes) necessary the software does the heavy job by typing a single command. What is most important is know the hard concepts and where to find the answers, and which test to be used in a certain situation or not. Table 8 briefly summarizes the most common tests and his usages.

Test Name	Usage
ANOVA (AN)	Tests to detect the mean difference of 3 or more independent groups.
Chi-Square (CH)	Tests the differences of the association between two categorical variables
Wilcoxon (TC)	Tests for the difference between two independent variables, takes into account magnitude and direction of difference
Fisher (FT)	Tests two nominal variables when is wanted to see whether the proportions of one variable are different depending on the value of the other variable. Use it when the sample size is small.
Kruskal-Wallis (KR)	Tests the medians of two or more samples to determine if the samples come from different populations.
Mann–Whitney (MW) Or U-Test (TU)	Test to compare two sample means that come from the same population, and used to test whether two sample means are equal or not.
T-Test (TT)	Tests for the difference between two related variables

Table 8. Statistical tests and his usage [9]

6.2 Follow a guideline

Although a good statistical analysis was done, as demonstrated, it is not described in most of the papers. This can occur for several reasons. Due to the fact that the researcher judge not necessary, due the small limit of pages of the conferences where these papers are submitted and the statistical analysis is deferred

in favor of the demonstration of the technique itself. Regardless the reason, the statistical analysis, as well as other stages of the research, must, even briefly, be described to demonstrate the robustness of the results and consequently of the technique developed.

In [13] is suggested that 85% of biomedical research effort is waste due to issues of low quality of results. Would be this the case of Computer Science too? It should be noted, on related works, that exist evidences, since 1993, that the results present in scientific works in Computer Science are weakly based from the statistical point of view. According to [21], the cause of this weak foundation is the little intimacy of the area with statistical methods (another reason to complement what was already described). However, these same works do not effectively present, or suggests, a verification method or guide to be followed before publication submission, and, similarly to the reviewers of these papers to establish selection criteria for publication in their journals.

In other areas of knowledge, such as the Medical for example, there are several protocols and guidelines to different types of studies developed in this area. These not only help during the development of the research, but also in the design phase as well as in the presentation of results including statistical analysis. In these documents are described, methodologically speaking, all sections of a scientific work and all particularities that need to be contained in each of them, according to the type of study performed.

According to [13], there are about 300 different types of medical research guides[2]. Initiatives such as **PRISMA**[3][11], **STROBE**[4][4][8], **CONSORT**[5][10] and **TOP**[6][16], are widely adopted by journals and periodicals as authors and reviewers guidelines. Individually, these initiatives do not meet the specifics of computing area, but could be a starting point to research guide and reporting, till a specifics computer science guidelines are not complete formulated.

In this way, initiatives of guidelines to fill this gap in computer science are still sparse. Most of them are restricted to the software engineering or simulation. When considered the verification of execution performance indexes, these initiatives do not address all possible types of studies or specificities such as the use of benchmarks and workloads as sample space [19]. This miss orientation, could contribute to impossibility do understand and reproduce most of studies reported in papers.

A literature review of this theme was done to investigate guidelines to research and reporting in Computer Science. The word "protocol" was discharged due to a bias related to network protocols. The research was limited to **papers** in **English** or **Portuguese**, published on the **last 10 years** (2009 to 2019). The terms "computer science", "reporting standards", "reporting evaluation", "reporting guidelines", "reporting statements", "reporting checklist", "research

[2] http://www.equator-network.org/library/
[3] Preferred Reporting Items for Systematic Reviews and Meta-Analyzes
[4] Strengthening the Reporting of Observational Studies in Epidemiology
[5] Consolidated Standards of Reporting Trials
[6] Transparency and Openness Promotion

standards", "research evaluation", "research guidelines", "research statements", "research checklist" were submitted to Springer Link [7], IEEE Digital Library[8], ACM Digital Library [9] and Science@Direct [10] obtaining a total of 1528 searched results. Where after paper title reading (62 selected), abstract reading (24 selected) and full paper reading, results in 10 papers with related information, but only four with complete relationship with the theme, the rest of them have some interest relation.

The highlight came from [6] showing as the most relevant reference of the 15 obtained. The related works cited by this publication in specific, was published before 2009. Thus the conclusion is that nothing new was published in the last decade in this sense, except the paper itself. If take into account the results of a literature review described above.

The concerns of [6] and [13] touched on the same situations raised here, reinforcing the argument. The lack of certain information in the publications may lead to the impossibility of understanding, replicating, auditing and evaluating the quality of the results obtained. In this respect [13] argues that the use of guidelines improves the orientation of researchers in both design and description of results.

Although it seem only bureaucratic processes, by the scientific community, guidelines is a robust way to improve the research as well as the reporting of it. [13] points that with the rapid adoption of guidelines in areas of knowledge, like Biomedical and Social Science, the problems related to the quality of the results of the researches are being overcome. Not only for individual researches, but for journals and periodicals. Publications that not adopted guidelines show poorer reporting quality compared with that have adopted[20] .

7 Conclusion

In this paper we present a research and analysis about the occurrence of statistical terms, metrics and statistical tests used to prove the results in scientific research. Publications of all WSCAD editions till 2017 were analyzed in a total sample of 426 papers.

From the analyzed sample, 398 publications referred to at least one of the searched terms, corresponding to 93% of the total. This shows that there is concern in research conducting and reporting, even inadequate or incomplete given the occurrence of only 3% of statistical tests confirmed by a second independent sample of 30 papers. It is necessary not only to focus on the design and development of the research object itself, but also on the measurements and reporting of the results.

Almost all of the papers only present the measurements of the implemented technique, and the questioning is inevitable: is just a simple measurement of

[7] http://link.springer.com
[8] http://ieeexplore.ieee.org
[9] http://portal.acm.org
[10] http://www.sciencedirect.com

the metric sufficient to compare similar techniques, disregarding error factors and considering that they were performed within the same standards? Are those results reproducible?

The research described here should be a warning to the scientific community. There is clearly a need for attention in the points highlighted. The findings of this work should promote reflection on undergraduate and postgraduate courses about the need to include methodology of statistical analysis of data applied to computing in basic formation classes. The adoption of research and reporting guidelines is a highly recommended alternative to mitigate issues, both with regard to study design, experimentation, collection of data of the experiment in adequate quantity and quality, in the application of statistical tests in a way that results have the effective power to show what is proposed.

The contribution with the WSCAD is to indicate in the call for papers the need for the papers to show their performance analysis accompanied by the methodological and statistical validation of their results, asking the reviewers of those papers to observe if such study was duly presented, by using specific and appropriate guide for the conference, which of course, needs to be formulated with the community of this event.

References

1. Adler, S., Schmitt, S., Wolter, K., Kyas, M.: A survey of experimental evaluation in indoor localization research. In: Indoor Positioning and Indoor Navigation (IPIN), 2015 International Conference on. pp. 1–10. IEEE (2015)
2. Bukh, P.N.D.: The art of computer systems performance analysis, techniques for experimental design, measurement, simulation and modeling (1992)
3. Dean, A., Voss, D., Draguljić, D., et al.: Design and analysis of experiments, vol. 1. Springer (1999)
4. Ebrahim, S., Clarke, M.: Strobe: new standards for reporting observational epidemiology, a chance to improve (2007)
5. Fortier, P., Michel, H.: Computer systems performance evaluation and prediction. Elsevier (2003)
6. de França, B.B.N., Travassos, G.H.: Experimentation with dynamic simulation models in software engineering: planning and reporting guidelines. Empirical Software Engineering **21**(3), 1302–1345 (2016)
7. Jain, R.: The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling. John Wiley & Sons (1990)
8. Malta, M., Cardoso, L.O., Bastos, F.I., Magnanini, M.M.F., Silva, C.M.F.P.d.: Iniciativa strobe: subsídios para a comunicação de estudos observacionais. Revista de Saúde Pública **44**, 559–565 (2010)
9. of Minnesota, U.: Types Of Statistical Testes (2019 (accessed May 23, 2019)), https://cyfar.org/types-statistical-tests
10. Moher, D., Hopewell, S., Schulz, K.F., Montori, V., Gøtzsche, P.C., Devereaux, P., Elbourne, D., Egger, M., Altman, D.G.: Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. International Journal of Surgery **10**(1), 28–55 (2012)

11. Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G.: Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. Annals of internal medicine **151**(4), 264–269 (2009)
12. Montgomery, D.C.: Design and analysis of experiments. John Wiley & Sons (2017)
13. Munafò, M.R., Nosek, B.A., Bishop, D.V., Button, K.S., Chambers, C.D., Du Sert, N.P., Simonsohn, U., Wagenmakers, E.J., Ware, J.J., Ioannidis, J.P.: A manifesto for reproducible science. Nature Human Behaviour **1**(1), 0021 (2017)
14. Navidi, W.: Probabilidade e estatística para ciências exatas. AMGH (2012)
15. Neto, B.B., Scarminio, I.S., Bruns, R.E.: Como Fazer Experimentos-: Pesquisa e Desenvolvimento na Ciência e na Indústria. Bookman (2010)
16. Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler, S.J., Buck, S., Chambers, C.D., Chin, G., Christensen, G., et al.: Promoting an open research culture. Science **348**(6242), 1422–1425 (2015)
17. Osorio, A., Dias, M., Cavalheiro, G.G.H.: Wscad: Uma meta-analise. In: WSCAD 2018 () (oct 2018), http://wscad.sbc.org.br/2018/anais/anais-wscad-2018.pdf
18. Prechelt, L.: A quantitative study of experimental evaluations of neural network learning algorithms: Current research practice. IEEE Transactions on Neural Networks **6** (1994)
19. Runeson, P., Höst, M.: Guidelines for conducting and reporting case study research in software engineering. Empirical software engineering **14**(2), 131 (2009)
20. Stevens, A., Shamseer, L., Weinstein, E., Yazdi, F., Turner, L., Thielman, J., Altman, D.G., Hirst, A., Hoey, J., Palepu, A., et al.: Relation of completeness of reporting of health research to journals' endorsement of reporting guidelines: systematic review. Bmj **348**, g3804 (2014)
21. Tedre, M., Moisseinen, N.: Experiments in computing: A survey. The Scientific World Journal **2014** (2014)
22. Tichy, W.F., Lukowicz, P., Prechelt, L., Heinz, E.A.: Experimental evaluation in computer science: A quantitative study. Journal of Systems and Software **28**(1), 9–18 (1995)
23. Wainer, J., Barsottini, C.G.N., Lacerda, D., de Marco, L.R.M.: Empirical evaluation in computer science research published by ACM. Information and Software Technology **51**(6), 1081–1085 (2009)
24. Wainer, J., et al.: Métodos de pesquisa quantitativa e qualitativa para a ciência da computação. Atualização em informática **1**, 221–262 (2007)