Context

Recently, Ali *et al.* presented a systematic review on how empirical investigations are conducted in search-based software testing.

The results concluded that the number of papers containing well-designed and well-reported empirical studies in the domain of SBST is very small.

In this research we extended Ali's review analyzing the papers published in the first two editions of the ISSBSE.

The goal is to evaluate how the usage of empirical studies in SBSE has improved in the last years.

At total, 23 papers were evaluated. We focused on the research questions presented in the next box.

Research Questions and Results

RQ1: *How well is the random variation inherent in SBSE accounted for in the design of empirical studies?*

		2009	2010	Total	%
Random Variation Accounted	Poor descriptive statistics	2	2	4	17%
	Good descriptive statistics	3	1	4	17%
	Statistical data analysis	2	7	9	40%
Random Variation not Accounted	Random variation not discussed	0	2	2	9%
	Insufficient number of runs	2	2	4	17%
TOTAL		**9**	**14**	**23**	**100%**

RQ2: *What are the most common search-based optimization algorithms used by SBSE techniques?*

Type of Search	Meta-heuristic Algorithms	2009	2010	Total	%
Global Search	Single-objective GA and extensions	2	9	11	46%
	Multiobjective GA and extensions	2	2	4	17%
	SA and extensions	2	0	2	8%
	Swarm based search	0	1	1	4%
Local Search	Hill Climbing	2	2	4	17%
	Tabu Search	1	1	2	8%
TOTAL		**9**	**15**	**24**	**100%**

RQ3: *What are the most common search-based alternatives to which SBSE techniques are compared?*

Type of Search	Meta-heuristic Algorithms	2009	2010	Total	%
Global Search	GA and extensions	1	8	9	25%
	SA and extensions	1	2	3	8%
	Swarm based search	0	1	1	3%
Local Search	Hill Climbing	2	4	6	16%
Non-Heuristic	Random Search	1	6	7	19%
	CBR	0	1	1	3%
	Expert	0	2	2	5%
	Regression	0	2	2	5%
	Other	3	1	4	11%
Not Discussed	-	2	0	2	5%

RQ4: *What are the measures used for assessing the cost of SBSE techniques?*

Cost Measure	2009	2010	Total	%
Execution time	2	6	8	35%
Iterations	2	0	2	9%
Fitness evaluations	5	3	8	35%
None	0	5	5	22%
TOTAL	**9**	**14**	**23**	**100%**

RQ5: *To which extent validity threats are addressed in reporting empirical studies in SBSE?*

Threat Group	Validity Threat Detail	LIMITED	PARTIAL	COMPLETE
Internal	Poor parameter settings	2 (9%)	0 (0%)	21 (91%)
	Lack of discussion on instrumentation of code	9 (39%)	5 (22%)	9 (39%)
	Lack of description of data collection procedures	4 (17%)	12 (52%)	7 (30%)
	Lack of using real problem instances	9 (39%)	9 (39%)	5 (22%)
Construct	Lack of validity of cost measures	10 (43%)	4 (17%)	9 (39%)
	Lack of validity of effectiveness measures	7 (30%)	8 (35%)	8 (35%)
	Lack of discussing the underlying model	2 (9%)	14 (61%)	7 (30%)
External	Lack of a definition of target instances	1 (4%)	10 (43%)	12 (52%)
	Lack of a object selection strategy	10 (43%)	0 (0%)	13 (57%)
	Lack of instances of growing size	9 (39%)	4 (17%)	10 (43%)
	Lack of instances of growing complexity	12 (52%)	7 (30%)	4 (17%)

Comparison to Ali et al.'s work

Ali et al. establishes that research papers showing credible evidences should …

1. account for the random variation in search algorithms and present a meaningful non-heuristic or local search comparison baseline (minimum criteria);

2. provide descriptive statistics on the variation of effectiveness and efficiency data collected after executing the selected algorithms or report on statistical inference tests applied to this data (sufficient criteria); and

3. address the scalability of the proposed algorithms in the context of the selected problem (scalability aware).

Below, we present the former classification for SBSE papers, comparing the numbers with those collected by Ali et al in their systematic review.

	Total papers	Minimum Criteria papers	Sufficient criteria papers	Scalability aware papers
Ali *et al.*	64	18 (28%)	8 (13%)	1 (2%)
SSBSE 2009	9	4 (44%)	2 (22%)	1 (11%)
SSBSE 2010	14	8 (57%)	8 (57%)	3 (21%)
SSBSE (all)	23	12 (52%)	10 (43%)	4 (17%)

Conclusions

We observed that the use of experimental studies in SBSE seems to have improved in the period between the two reviews.

As future work, we intend to perform a regular systematic review upon SBSE papers from different sources, using the same framework with the purpose of confirming the observed trends.

PPGI - UNIRIO