Do Quality Indicators Prefer Particular Multi-Objective Search Algorithms in Search-Based Software Engineering?*

Shaukat Ali[1 [0000−0002−9979−3519]], Paolo Arcaini[2 [0000−0002−6253−4062]], and Tao Yue[1,3 [0000−0003−3262−5577]]

[1] Simula Research Laboratory, Oslo, Norway
[2] National Institute of Informatics, Tokyo, Japan
[3] Nanjing University of Aeronautics and Astronautics, Nanjing, China

Abstract. In Search-Based Software Engineering (SBSE), users typically select a set of Multi-Objective Search Algorithms (MOSAs) for their experiments without any justification, or they simply choose an MOSA because of its popularity (e.g., NSGA-II). On the other hand, users know certain characteristics of solutions they are interested in. Such characteristics are typically measured with Quality Indicators (QIs) that are commonly used to evaluate the quality of solutions produced by an MOSA. Consequently, these QIs are often employed to empirically evaluate a set of MOSAs for a particular search problem to find the best MOSA. Thus, to guide SBSE users in choosing an MOSA that represents the solutions measured by a specific QI they are interested in, we present an empirical evaluation with a set of SBSE problems to study the relationships among commonly used QIs and MOSAs in SBSE. Our aim, by studying such relationships, is to identify whether there are certain characteristics of a QI because of which it prefers a certain MOSA. Such preferences are then used to provide insights and suggestions to SBSE users in selecting an MOSA, given that they know which quality aspects of solutions they are looking for.

Keywords: Search-Based Software Engineering · Quality Indicator · Multi-Objective Search Algorithm

1 Introduction

Researchers and practitioners (i.e., users) in Search-Based Software Engineering (SBSE) dealing with multi-objective problems, need to select multi-objective search algorithms (MOSAs) to solve them. However, the use of a specific MOSA in a particular SBSE context is rarely justified [14]. At the same time, users are aware of certain characteristics of solutions that they are interested to find with MOSAs. Such characteristics are measured with Quality Indicators (QIs) that are commonly employed to assess the quality of solutions produced by MOSAs from different perspectives. Consequently, these

* This work is supported by the Co-evolver project (No. 286898/F20) funded by the Research Council of Norway, National Natural Science Foundation of China under Grant No. 61872182, and ERATO HASUO Metamathematics for Systems Design Project (No. JPMJER1603), JST, Funding Reference number: 10.13039/501100009024 ERATO.

QIs are often used by users to select the best MOSA for their particular applications. To this end, the motivation of this paper is to help SBSE users in selecting an MOSA, i.e., if they are interested in the qualities of solutions preferred by a given QI, we identify the MOSA that will likely produce solutions having these qualities.

This application context is common in practical applications, where one knows which qualities in solutions they prefer and only wants to select one suitable MOSA without conducting extensive experiments to compare multiple MOSAs, which are often expensive. Indeed, in practical contexts, users have limited time budget to run experiments involving multiple MOSAs and would like to select one that produces the "best" solutions in terms of preferred QIs. Thus, in this paper, we aim to provide evidence that is useful for users to choose an MOSA that is highly likely to give solutions satisfying qualities measured by a preferred QI.

There exist surveys and studies about investigating various quality aspects of QIs and their relationships with MOSAs in the context of SBSE. For example, the survey of Sayyad et al. [14] reports that most of the publications in SBSE that were included in the survey do not provide justifications on why one or more particular MOSAs have been chosen in their experiments. The survey also reports that, in some of the investigated publications, researchers compared MOSAs against each other with certain QIs. This indicates that researchers were aware of the quality aspects provided by the selected QIs, and were looking for MOSAs that can produce solutions entailing such quality aspects. The survey also concluded that researchers sometimes chose MOSAs only based on their "popularity", i.e., selecting commonly used MOSA(s).

In the literature, there are also studies on investigating relationships of QIs and their characteristics. For instance, in our previous paper [1], we analyzed agreements among QIs commonly used in SBSE, with the aim to provide users with a set of guidelines on selecting QIs for their SBSE applications. Similarly, Li and Yao [10] surveyed 100 QIs that have been used in the evolutionary computation domain with the aim of studying their strengths and weaknesses. In the current SBSE literature, however, relationships between QIs and MOSAs are not well studied. In particular, the question of whether there are certain MOSAs that are preferred by a given QI is not answered. Answering such research question can help in guiding users to select an MOSA, on the basis of their preferences of QIs.

In this paper, we present an empirical evaluation in SBSE to study relationships between the QIs and the MOSAs to provide evidence indicating which MOSA(s) are highly likely to produce solutions that entail the given quality aspects represented by a specific QI. The empirical evaluation was performed with various industrial, real-world, and open source SBSE problems available online with commonly used MOSAs and QIs in SBSE. Our results reveal that certain QIs prefer specific MOSAs (e.g., Hypervolume prefers NSGA-II), whereas some QIs (e.g., the ones involving the quality aspect of *Cardinality* [10]) do not have any strong preference. Based on our results, we present a set of suggestions and insights to select an MOSA based on a particular QI or a category of quality aspects. The rest of the paper is organized as follows: Section 2 relates our work with the existing works in the literature. Section 3 shows the design of our empirical evaluation, and Sections 4 and 5 describe analyses and results. Section 6 presents

the discussion and our recommendations, whereas threats to validity are presented in Section 7. Finally, Section 8 concludes the paper.

2 Related Work

Sayyad and Ammar [14] presented a survey on SBSE papers that use MOSAs for solving software engineering optimization problems, from the perspectives of the chosen algorithms, QIs, and used tools. The paper concludes that more than half of the 51 surveyed papers do not provide justifications on the selection of a specific MOSA for a specific problem or simply state that an MOSA is selected because it is often applied by others. This observation, to a certain extent, implies that in the SBSE research community, there is no evidence showing which MOSA(s) to apply, in particular in the context in which researchers do know which QI(s) they prefer. Our current study provides evidence for guiding researchers selecting MOSAs based on the preferences of QIs.

The most relevant work, though not in the SBSE context, was presented by Ravber et al. in [13]. The work studied the impact of 11 QIs on the rating of 5 MOSAs: IBEA, MOEA/D, NSGA-II, PESA-II, and SPEA2, and concluded that QIs even with the same optimization goals (*convergence*, *uniformity*, and/or *spread*) might generate different and contradictory results in terms of ranking MOSAs. The authors analyzed the 11 QIs using a Chess Rating System for Evolutionary Algorithms [16], with 10 synthetic benchmark problems from the literature and 3 systems for a real-world problem. Based on the results of the analysis, the studied QIs were categorized into groups that had insignificant differences in ranking MOSAs. A set of guidelines were briefly discussed, considering preferred optimization aspects (e.g., *convergence*) when selecting QIs for a given search problem and selecting a robust (achieving the same rankings of MOSAs for different problems) and big enough set of QIs. To compare with our work reported in this paper, our study differentiates itself from [13] in the following two aspects. First, our study focuses exclusively on SBSE problems, whereas their study was conducted in a general context, and therefore the sets of MOSAs and QIs used in the two studies are different. The MOSAs and QIs we selected in our study are commonly applied ones in the context of SBSE. Second, our study aims to provide evidence on selecting an MOSA for solving an SBSE problem, in the context in which the user is aware of the desired quality aspects in the final solutions, and has limited time budget (in terms of running experiments). Instead, their study aims to suggest which QI(s) to select for assessing MOSAs. When looking at the results of both studies on ranking MOSAs for each QI, there are similarities and dissimilarities, details of which will be discussed in Section 5.

Li and Yao reported a survey [10] on 100 QIs from the literature, discussed their strengths and weaknesses, and presented application scenarios for a set of QIs. In this survey, only two studies [18,9] related to SBSE were included, which are about understanding QIs from various aspects. Wang et al. [18] proposed a guide for selecting QIs in SBSE based on the results of an experiment with 8 QIs, 6 MOSAs, and 3 industrial and real-world problems. Their guide helps to determine a category of the QIs (*Convergence*, *Diversity*, *Combination of convergence and diversity*, or *Coverage*). In our previous work [1], we conducted an extensive empirical evaluation with 11 SBSE search problems from industry, real-world ones, and open source ones, and automat-

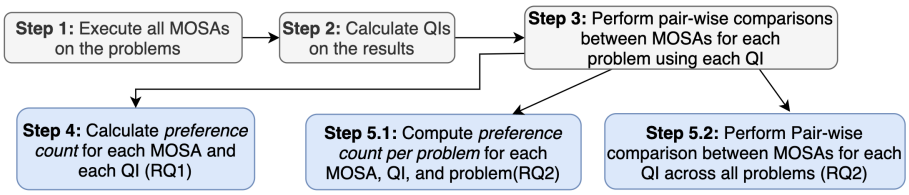


Fig. 1: Design of the Experiment

ically produced 22 observations based on the results of the statistical tests for studying QI agreements, by considering different ways in which SBSE researchers typically compare MOSAs. We also provided a set of guidelines in the form of a process that can be used by SBSE researchers. To compare with our previous work [1], in this paper, we aim to suggest which MOSA to select given a QI that is preferred, while previously, we aimed at suggesting which QI(s) to use for evaluating a given MOSA.

3 Design of Empirical Evaluation

The design of our empirical evaluation is shown in Fig. 1. It consists of five steps. Steps 1-3 were performed by the authors of the papers [1,17], who afterwards made the data publicly available. We used this public data for our empirical evaluation and performed Steps 4-5. These steps will be described in following subsections.

3.1 Selection of Search Problems, MOSAs, and QIs

For our empirical evaluation, we chose data from SBSE search problems that are described in details in [1]. The data consists of a mix of industrial, real-world, and open source SBSE problems. In total, there were data available for 11 SBSE search problems. The data also had the results of 100 runs of commonly used MOSAs in SBSE (Step 1 in Fig. 1), i.e., NSGA-II [5], MoCell [12], SPEA2 [20], PAES [8], SMPSO [11], CellDE [7], and Random Search (RS). CellDE was not applicable to one of the SBSE problems that requires Integer solutions, because CellDE works for Real type solutions only [6]. The chosen MOSAs were run using appropriate parameter settings of the MOSAs based on the previous experiments provided in [1]. These parameter settings can be found in in [1]. In addition, the available data also had computed QIs for the commonly used QIs in SBSE (Step 2 in Fig. 1), i.e., Generational Distance (GD), Euclidean Distance (ED), Epsilon (EP), Generalized Spread (GS), Pareto Front Size (PFS), Inverted Generational Distance (IGD), Hypervolume (HV), and Coverage (C). The definition of these QIs can be consulted in [10]. Moreover, the authors performed relevant statistical tests to compare each pair of MOSAs using each QI (Step 3 in Fig. 1). The results of the tests reveal which MOSA performed significantly better than the other one with respect to a particular QI. Note that all the MOSAs performed significantly better than RS; therefore, we did not include the results of RS. If we include the results of RS in our experiments, it will always be the least preferred by all the QIs.

The results of these statistical tests were used in our empirical evaluation reported in this paper: namely, we used them to perform Step 4 and Step 5, in order to answer the RQs defined in the next subsection.

3.2 Research Questions

Our overall objective is to study the relationships between the QIs and the MOSAs with the aim of finding whether there are specific characteristics of MOSAs that are preferred by a specific QI. To this end, we define the following Research Questions (RQs):

- RQ1: How frequently a QI prefers a particular MOSA? This RQ studies the percentage of times that a QI prefers a particular MOSA by ignoring the differences of the SBSE problems when studying pairs of MOSAs. This RQ helps in understanding the overall preferences of a QI.
- RQ2: How frequently a QI prefers a particular MOSA across the different SBSE problems? This RQ studies the preferences of QIs across the problems when studying pairs of MOSAs, whereas in RQ1 we aim to study preferences while ignoring the differences of the problems.

3.3 Evaluation Metrics, Statistical Analyses, and Parameter Settings

We define a set of evaluation metrics to answer the two RQs. First of all, we use the predicate $pref$ to indicate the preference relationship between MOSAs for a given quality indicator and a given problem. Let Q be a quality indicator, A and B two MOSAs, and P a search problem. $pref(A, B, Q, P) = true$ iff Q prefers MOSA A to B when these are applied to the search problem P. The preference relation has been computed in our previous work [1], where we compared 100 runs of MOSA A and MOSA B over problem P using the *Mann-Whitney U* test and *Vargha and Delaney* $\hat{A}_{12}$ statistic. Note that $pref(A, B, Q, P) = true$ implies $pref(B, A, Q, P) = false$. If $pref(A, B, Q, P) = false$ and $pref(B, A, Q, P) = false$, it means that Q does not have any significant preference among the two MOSAs.

RQ1 In order to answer RQ1 (*Step 4* in Fig. 1), we introduce the following measure. Let $MOSAs$ be the set of MOSAs, $Problems$ the set of search problems, and Q a quality indicator. We define the *preference count* as the percentage of times Q prefers MOSA A when compared to another MOSA in any problem, formally:

$$PC(A, Q) = \frac{|\cup_{P \in Problems} \{B \in (MOSAs \setminus \{A\}) \mid pref(A, B, Q, P)\}|}{(|MOSAs| - 1) \times |Problems|} \quad (1)$$

The rationale is that if an MOSA A is consistently preferred by a QI Q (when compared with other MOSAs and for different problems), it means that A tends to produce solutions that have the quality aspects assessed by Q. The higher $PC(A, Q)$ is, the higher the probability is that, also on new problems, A will produce solutions preferred by Q.

RQ2 In order to answer RQ2, first, we compute the *preference count per problem* defined as follows (*Step 5.1* in Fig. 1)[4]:

$$PC(A, Q, P) = \frac{|\{B \in (MOSAs \setminus \{A\}) \mid pref(A, B, Q, P)\}|}{|MOSAs| - 1} \quad (2)$$

Second, we also perform, for each Q, pair-wise comparisons of the selected MOSAs across search problems (*Step 5.2* in Fig. 1). We choose the *Mann-Whitney U* test to determine the statistical significance of results, whereas we choose the *Vargha and Delaney* $\hat{A}_{12}$ statistics as the effect size measure. These statistical tests were chosen based on the published guidelines reported in [3]. When comparing two algorithms A and B with respect to a QI Q, if the p-value computed by the Mann-Whitney U test is less than 0.05 and $\hat{A}_{12}$ is greater than 0.5, then it means that A is significantly better than B with respect to Q. Similarly, when a p-value is less than 0.05 and $\hat{A}_{12}$ is less than 0.5, it means that B is significantly better than A. Finally, a p-value greater than or equal to 0.05 implies no significant differences between A and B with respect to Q.

The results of these tests give a more trustworthy definition of preference between MOSAs. In order to distinguish it from the one used in RQ1, we will call it *significant preference*, i.e., we will say that MOSA A is *significantly preferred* over MOSA B. Note that in Step 5.1 of Fig. 1, we study preferences per problem, whereas in Step 5.2 we study preferences across the problems. Moreover, in Step 4 we count preferences, while in RQ2 we determine the significance of preferences with the statistical tests.

4 Results and Analyses

In this section, we present the results and analyses for our RQs. Section 4.1 presents the results of RQ1, whereas Section 4.2 presents the results of RQ2.

4.1 RQ1

Recall that RQ1 aims to study the percentage of times that a QI Q prefers a particular MOSA A across all the problems when comparing pairs of MOSAs using the relevant statistical tests (see Eq. 1 in Section 3.3). Answering this RQ helps us in understanding the overall preferences of QIs.

The results are reported in Table 1. The *Preferred (%)* columns show the percentages calculated with the formula $PC(A, Q)$ presented in Section 3.3 (Eq. 1). A percentage determines the *preference count* of a QI for each MOSA across all the SBSE problems.

Based on the results, we can see that some QIs seem to prefer particular MOSAs. For example, HV prefers NSGA-II (75.93%) and SPEA2 is preferred by GD the most (77.78%). This observation suggests that NSGA-II may have some characteristics that are preferred by HV, and SPEA2 has some characteristics that are preferred by GD.

[4] Note that CellDE is not applicable to one of the search problems, and so the formulations of Eq. 1 and Eq. 2 should be slightly more complicated. We report the simplified versions here, but we use the correct versions in the experiments.

Table 1: RQ1 – *Preference count*

QI	MOSA	Preferred (%)	QI	MOSA	Preferred (%)	QI	MOSA	Preferred (%)	QI	MOSA	Preferred (%)
HV	NSGA-II	75.93%	EP	NSGA-II	74.07%	GS	SMPSO	74.07%	PFS	NSGA-II	62.96%
HV	SPEA2	68.52%	EP	SPEA2	70.37%	GS	CELLDE	64%	PFS	SPEA2	57.41%
HV	SMPSO	55.56%	EP	SMPSO	62.96%	GS	SPEA2	61.11%	PFS	SMPSO	35.19%
HV	CELLDE	42%	EP	CELLDE	46%	GS	NSGA-II	57.41%	PFS	PAES	31.48%
HV	MOCELL	29.63%	EP	MOCELL	27.78%	GS	MOCELL	22.22%	PFS	CELLDE	28%
HV	PAES	7.41%	EP	PAES	5.56%	GS	PAES	5.56%	PFS	MOCELL	18.52%
IGD	NSGA-II	77.78%	GD	SPEA2	77.78%	ED	SPEA2	66.67%	C	SPEA2	46.3%
IGD	SPEA2	77.78%	GD	NSGA-II	75.93%	ED	NSGA-II	62.96%	C	NSGA-II	38.89%
IGD	SMPSO	46.3%	GD	MOCELL	42.59%	ED	SMPSO	44.44%	C	CELLDE	22%
IGD	MOCELL	35.19%	GD	CELLDE	36%	ED	CELLDE	42%	C	MOCELL	18.52%
IGD	CELLDE	34%	GD	SMPSO	33.33%	ED	MOCELL	33.33%	C	SMPSO	16.67%
IGD	PAES	16.67%	GD	PAES	24.07%	ED	PAES	22.22%	C	PAES	7.41%

Table 2: RQ1 – Overall ranking of MOSAs

Rank	Instances for each MOSA
1	NSGA-II (3), SPEA2 (3), NSGA-II/SPEA2 (1), SMPSO (1)
2	NSGA-II (3), SPEA2 (3), CELLDE (1)
3	SMPSO (4), MOCELL (2), SPEA2 (1), CELLDE (1)
4	CELLDE (5), NSGA-II (1), PAES (1), MOCELL (1)
5	MOCELL (4), SMPSO (3), CELLDE (1)
6	PAES (7), MOCELL (1)

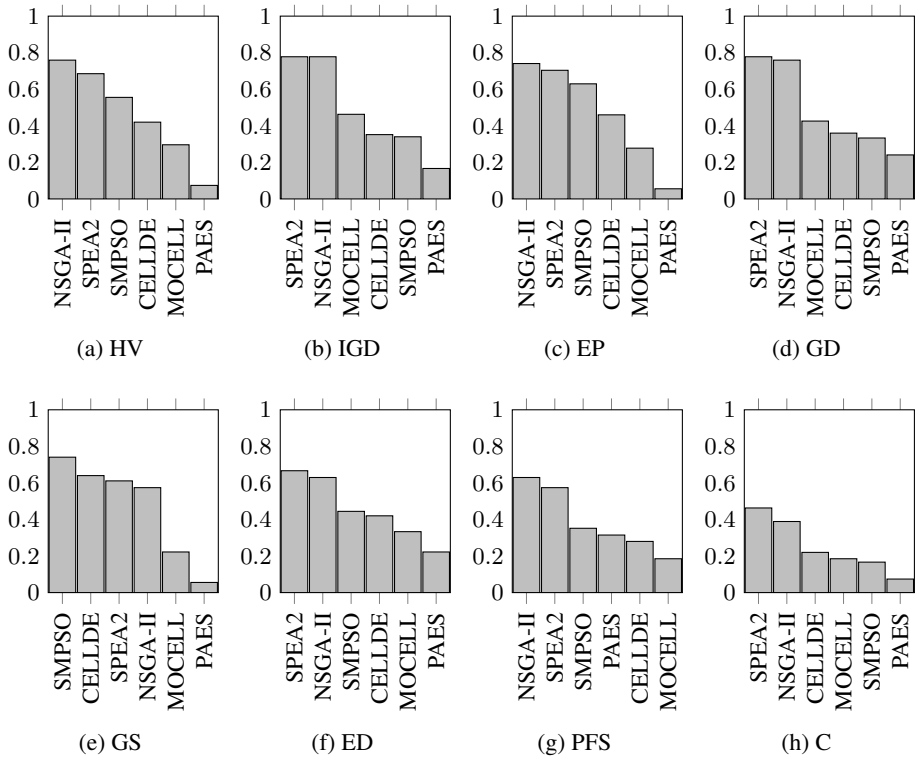
From the table, we can also observe that some QIs have low preference for some MOSAs, e.g., EP with PAES (5.56%). This means that such MOSAs do not usually produce solutions that have the qualities preferred by these QIs. Some QIs don't have strong preferences for any MOSA. For example, C has low percentages for all the MOSAs, thus suggesting that C assesses qualities that are not peculiar of any MOSA.

Moreover, in Fig. 2, we present preferences of each QI for all the selected MOSAs, sorted based on the percentages. For instance, by looking at HV and EP, we can see that NSGA-II was the most preferred by HV and EP, followed by SPEA2, whereas PAES was the least preferred by HV and EP. Similarly, GD, ED, and C prefer SPEA2 the most, followed by NSGA-II, and they least prefer PAES.

We provide a summary of the results in terms of which MOSAs are most preferred across all the QIs in Table 2. We can see from the table that NSGA-II and SPEA2 are on the top rankings (i.e., 1 and 2), meaning that these two MOSAs are the most preferred ones by most of the QIs, except for the cases of SMPSO for GS for rank 1 and CellDE for rank 2. We can also observe that PAES is least preferred by seven out of eight studied QIs. For PFS, MOCELL was the least preferred.

4.2 RQ2

Fig. 3 reports, for each quality indicator Q and each MOSA A, the distribution of the metric *preference count per problem* $PC(A, Q, P)$ across search problems P (see



(a) HV (b) IGD (c) EP (d) GD

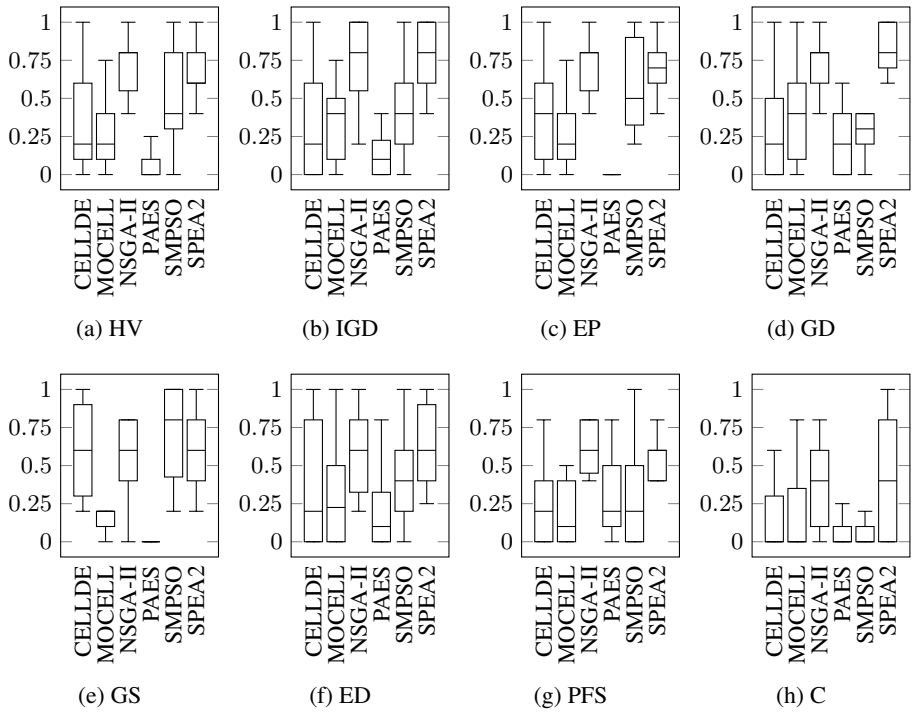
(e) GS (f) ED (g) PFS (h) C

Fig. 2: RQ1 – MOSAs sorted by *preference count*

Eq. 2). Of course, for MOSAs having a general low *preference count* (see Table 3), the variance of metric *preference count per problem* across problems is low. On the other hand, for MOSAs that in general are preferred, the variance is higher. This means that problem characteristics can influence the effectiveness of a particular MOSA A and so some quality indicator Q may prefer MOSA A on some problems, but not on some others. Note that the influence of the problem characteristics on the results of QIs has been discovered in a different setting. In our previous work [1], we discovered that the agreement between pairs of QIs, i.e., whether they prefer the same MOSA, sometimes depends on problems solved by MOSAs.

Note that we cannot perform an analysis on the base of the different problem characteristics (e.g., number of objectives), as this would require many more problems to have enough problems for each given characteristic. For the number of objectives, for example, we have four problems with two objectives, three problems with three objectives, and four problems with four objectives. This is not enough to draw any conclusion about the influence of the number of objectives.

Table 3 reports the overall results from the statistical test we performed. Recall from Section 3.3 that we performed the Mann-Whitney U test and the $\hat{A}_{12}$ statistics. A



(a) HV (b) IGD (c) EP (d) GD

(e) GS (f) ED (g) PFS (h) C

Fig. 3: RQ2 – Distribution of *preference count per problem* over search problems

Table 3: RQ2 – Overall preferences of QIs

	MOSA					
	CELLDE	MOCELL	NSGA-II	SMPSO	SPEA2	PAES
C	0	0	2	0	2	0
ED	0	0	2	0	2	0
EP	1	1	2	2	2	0
GD	0	0	4	0	4	0
GS	2	1	2	2	2	0
HV	1	1	3	1	2	0
IGD	0	0	4	1	4	0
PFS	0	0	4	0	3	0

number in a cell of the table means the number of times that an MOSA (e.g., NSGA-II) was *significantly preferred* over other MOSAs, i.e., a p-value with the U Test was less than 0.05 and the $\hat{A}_{12}$ value greater than 0.5. For example, with respect to GD, NSGA-II was significantly preferred over other four MOSAs.

Based on the results, we can see that, for all the QIs, NSGA-II is significantly preferred by all the selected QIs, since the columns for NSGA-II has either the higher

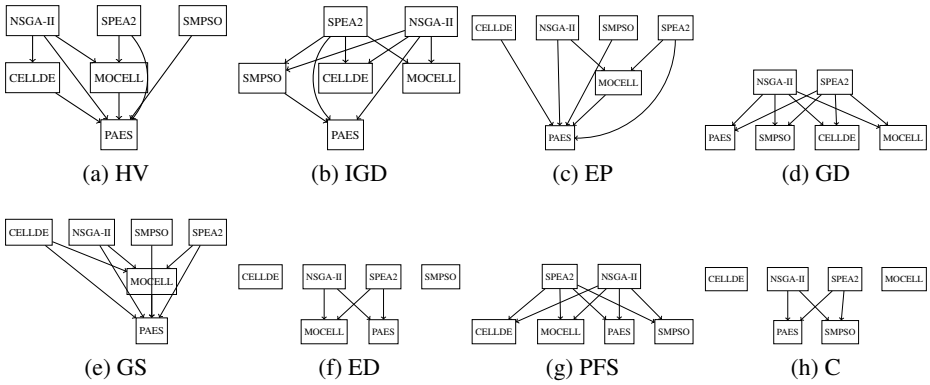


Fig. 4: RQ2 – Significant preference relation between MOSAs pairs

numbers (e.g., PFS) or equal numbers (e.g., EP) as compared to other MOSAs. Followed by NSGA-II, SPEA2 is the most preferred, having all the results the same as NSGA-II except for PFS, where NSGA-II has value 4, whereas SPEA2 has value 3.

Fig. 4 shows a more detailed representation of the significant preference relation. For each QI, it shows which MOSAs are significantly preferred over others. An arrow from MOSA A to MOSA B means that A is significantly preferred over B. We observe that some MOSAs are consistently significantly preferred over some others, as NSGA-II and SPEA2 that are always preferred over PAES. The most preferred MOSAs (i.e., those with the highest numbers in Table 3) are usually preferred over the same other MOSAs. Moreover, there are some MOSAs that, although are worst than some MOSAs, are better than some others. For example, MOCELL in HV, EP, and GS. These are MOSAs that, although cannot produce optimal solutions, can still produce good ones.

5 Analyses based on Quality Aspects of QIs

At a higher level, each QI covers certain *quality aspects* of solutions produced by an MOSA. Thus, we analyzed our results from a different perspective, i.e., we checked whether an MOSA is preferred by QIs that cover certain quality aspects.

To this end, we present results of overall preferences for various categories of quality aspects measured by QIs. Li and Yao [10] categorized such quality aspects into four categories. The first category is *Convergence* that focuses on measuring the quality of a set of solutions on a Pareto front based on how close these solutions are to a reference Pareto front. The second category is *Spread*, which measures the quality by measuring the spread of solutions. The third category is *Uniformity* that measures the quality by looking at the distribution of solutions in terms of how evenly these solutions are distributed. Finally, the *Cardinality* category focuses on counting the number of solutions on a Pareto front. We chose this classification since it is published in the recent survey by Li and Yao [10] that studied 100 QIs from the literature. Table 4 shows which of our selected QIs fully or partially cover which categories of quality aspects.

Table 4: Relation between QIs and categories of quality aspects [10]

	QI							
	C	ED	EP	GD	GS	HV	IGD	PFS
Convergence	-	-	+	+		+	+	
Spread			+		-	+	+	
Uniformity			+		+	+	+	+
Cardinality			-			-	-	+

A cell with a "+" signifies that a particular quality aspect is fully represented by a QI, where "-" signifies a partial representation.

Table 5: Overall preferences of QI Categories

	MOSA					
	CELLDE	MOCELL	NSGA-II	SMPSO	SPEA2	PAES
Convergence	2	2	17	4	16	0
Spread	4	3	11	6	10	0
Uniformity	4	3	11	6	10	0
Cardinality	2	2	13	4	11	0

Table 5 summarizes our results. For each MOSA A and each category, it reports how many times A is preferred (as reported in Table 3) by a QI belonging to the considered category. For example, for the *Convergence* category, NSGA-II has the highest value, i.e., 17. In general, we can see that NSGA-II is the most preferred one in all the four categories. However, note that these results are based on a particular set of SBSE problems, and once additional SBSE problems are added, these results may change.

We also checked whether QIs that cover the same quality aspects have also the same preferences for MOSAs. Looking at Table 4, we can see that C and ED are partially represented by the *Convergence* quality aspect. Now, looking at the results of C and ED in Fig. 2, we can see partial similarity, i.e., they both preferred SPEA2 and NSGA-II at the first and second place, whereas they both least preferred PAES. However, the other three MOSAs were in different positions. In addition, we can see that for C the preference count is in general lower for all the MOSAs as compared to ED.

As a further observation, we note that the quality indicators EP, HV, and IGD fully represent *Convergence*, *Spread* and *Uniformity*, and partially represent *Cardinality* as shown in Table 4. For HV and EP, we observe the same results in Fig. 2. However, for HV/EP and IGD, we see some differences. For example, SPEA2 and NSGA-II are equally at the first place for IGD, whereas NSGA-II is at the first place followed by SPEA2 for HV/EP. Also, MOCELL, CELLDE, and SMPSO are at different places.

Based on these observations, we can conclude that the QIs covering the same quality aspects don't necessarily have the same preferences for MOSAs. This observation is consistent with what has been reported in [13], in which the authors concluded that QIs with the same quality aspect(s) do not necessarily yield the same rankings of MOSAs.

6 Overall Discussion

Here, we present an overall discussion both for the results based on individual QIs and for the QI categories. Moreover, we provide suggestions to users for selecting an MOSA that will likely produce solutions preferred by a given QI or a given quality aspect.

When we look at the results based on QIs (see Table 3), we observe that both NSGA-II and SPEA2 are highly preferred by each QI. However, one must understand that some results provide more confidence than others when we want to suggest to pick a specific MOSA on the base of these results. For example, for HV, NSGA-II has a value of 3 out of 5 (i.e., 60% of the times NSGA-II was preferred over the other MOSAs), whereas, for PFS, NSGA-II has a value of 4 out of 5 (i.e., 80%). In both cases in which a user wants solutions that are represented by HV or PFS, our results suggest to use NSGA-II. However, in the latter case, the user will be more confident to follow the suggestion, because this is based on a stronger result.

Sometimes there is a tie among the MOSAs. For example, for EP, NSGA-II, SMPSO, and SPEA2 all have a value of 2. In these situations, we suggest the following options:

(1) selecting any MOSA, or
(2) selecting the MOSA from the category table (i.e., Table 5) by checking which quality aspect(s) is(are) represented by the selected QI (i.e., Table 4). For example, suppose that a user selected C. From Table 3, we see that there is a tie between NSGA-II and SPEA2 for C (value 2). In this case, the user may consult Table 4 and find that C represents solutions with the *Convergence* quality aspect. Then, the user can see from Table 5 that, for *Convergence*, NSGA-II is the preferred MOSA. Note that a QI could represent more than one category of quality aspects (e.g., HV). In our context, the preferred MOSA is always NSGA-II. However, when more data are available, the preferred MOSAs may be different for the different categories of quality aspects. In this case, more complicated guidelines could be provided (e.g., taking the MOSA scoring the highest in the majority of categories), or the user could decide to select the MOSA associated with the quality aspect they prefer.

Now, looking at our results for QI categories (Table 5), we note that NSGA-II is the clear option; however, once the results are updated based on additional experimental results of other problems, the preferences may change and we may have ties between two or more MOSAs. If such a case arises, we suggest selecting any of the preferred MOSAs of the tie or taking into account other aspects of MOSAs (e.g., their time performance).

Note that, as also observed in Section 4.2, in our experiments we did not study relationships between QI preferences and characteristics of the SBSE problems (e.g., search objective types, data distributions). Such characteristics could help us to provide a better guidance for SBSE users based on different characteristics of SBSE problems. Please note that conducting such an experiment requires a complete and well-planned experiment of its own, involving controlling various characteristics of SBSE problems in a systematic way. Finding publicly available SBSE problems that systematically cover various characteristics is challenging and one may resort to creating synthetic problems. We plan to conduct such an experiment in the future, where we could also study characteristics of QIs and SBSE problems together to suggest appropriate MOSAs.

7 Threats to Validity

We here discuss threats that may affect the validity of our experiments, namely *internal validity*, *conclusion validity*, *construct validity*, and *external validity* [19].

Internal validity: Many MOSAs have been proposed in the literature, and a threat to the internal validity is that we may have not considered an MOSA that is consistently better than those considered in this paper. In order to address such a threat, we selected the MOSAs that are commonly used in SBSE [17,1].

Another threat is related to the settings of the parameters of the selected MOSAs. an MOSA A may perform better (i.e., preferred by a given QI) than another MOSA B, because it has been configured better. In order to address such threat, we have configured the selected MOSAs by following the commonly applied guides [3,15]. Note that these same settings were used in the papers from which we obtained the case studies, and in those papers these settings have been proven to give good results.

In terms of selection of QIs, one may argue that we did not cover enough QIs, given that there exist 100's of QIs [10]. However, note that we selected the most commonly used QIs in the SBSE literature [14], since our empirical evaluation was focused on SBSE problems. When presenting our results based on QI categories, one may wonder why we did not choose other QI categories, such as the ones proposed by Wang et al. [17]. We chose the QI categories instead from a recent survey [10], which is based on the study of 100 QIs, since it performs extensive evaluation of the existing QIs. Moreover, the QI categories in Wang et al. [17] are not precise as argued in [9].

Finally, we would like to mention that QIs may have different preferences thresholds to determine the significance of preferences. Such thresholds weren't studied in this paper and will require additional experiments.

Conclusion validity: One such a threat is that the input data that we used in our experiment may not be sufficient to draw conclusions between the application of an MOSA A on a given problem P, and its evaluation with a given QI. To mitigate such a threat, we have selected benchmarks in which each MOSA has been run 100 times, in order to reduce the effect of random variations. The conclusion whether a QI prefers an MOSA A to an MOSA B for a given problem P, is decided using the Mann-Whitney U test over the distribution of 100 QI values for A and B. Note that, in order to mitigate another threat related to wrong assumption for the tests, we selected such tests by following guidelines for conducting experiments in SBSE [3].

Construct validity: One construct validity threat is that the measures we used for drawing our conclusions may not be adequate. As first measure, we computed the percentage of times that an MOSA A is preferred over another MOSA B by a given QI Q, since our aim is to suggest an MOSA that will likely produce solutions preferred by Q. Hence, we believe that this metric is adequate. Moreover, to draw more stable conclusions, we also assessed the statistical significance of the results with the Mann-Whitney U test and the $\hat{A}_{12}$ statistics. More specifically, we compared the *preference counts per problem* (see Eq. 2) of the two MOSAs across the problems with the statistical tests for each QI. Note that, in rare cases, the p-value with the Mann-Whitney U test may be less than 0.05, but the $\hat{A}_{12}$ gives a value still close to 0.5. This means that differences are not representing an actual preference. We need to look into such cases more carefully in the future.

External validity: A major threat is that the results may not be generalizable to other case studies. In order to address such a threat, we selected as many SBSE problems as possible and ended up with 11 problems in total, trying to cover different types of SBSE problems: rule mining in product line engineering, test optimization, and requirements engineering. However, we are aware that such a selection is inherently partial, and we need more case studies from more SBSE problems to generalize the results. The lack of real-world case studies to be used in empirical studies is recognized to be a common threat to external validity [2,4]. Note that the work presented in this paper does not aim at giving ultimate results, but at providing a methodology that should be followed to build a body of knowledge about the relationship between MOSAs and QIs. To this aim, we make our implementation publicly available[5] and invite SBSE researchers to share with us their empirical studies, so to derive more reliable conclusions.

8 Conclusion and Future Work

We were motivated by the observation that, in the community of search-based software engineering (SBSE), users (researchers and practitioners) often need to select a multi-objective search algorithm (MOSA) for their application, especially in the situation that the users do not have sufficient time budget to conduct experiments to compare multiple MOSAs. Though in the literature there exist works studying QIs (their characteristics and relationships), the relationships between QIs and MOSAs are however not sufficiently studied. Motivated by this, in this paper, we presented an empirical evaluation and provided evidence to help users to choose an MOSA that is highly likely to produce solutions satisfying qualities measured by a QI preferred by the users. Specifically, we observed that NSGA-II and SPEA2 are preferred by most of the QIs we investigated; PAES is not preferred by most of the QIs. However, we would like to point out that, when selecting an MOSA, in addition to the quality aspects covered in each QI, other aspects such as time performance of MOSA should be considered as well. In the future, we would like to include such aspects in our study.

In addition, we would also like to design experiments for studying each specific quality aspect (e.g., Convergence). Furthermore, when more data will be available, we will conduct more analyses and update our findings. Finally, we would also like to study the preferences of QIs together with various characteristics of search problems. Such study can help users to select MOSAs based on characteristics of SBSE problems.

References

1. Ali, S., Arcaini, P., Pradhan, D., Safdar, S.A., Yue, T.: Quality indicators in search-based software engineering: An empirical evaluation. ACM Trans. Softw. Eng. Methodol. **29**(2) (Mar 2020). https://doi.org/10.1145/3375636
2. Ali, S., Briand, L.C., Hemmati, H., Panesar-Walawege, R.K.: A systematic review of the application and empirical investigation of search-based test case generation. IEEE Trans. Softw. Eng. **36**(6), 742–762 (Nov 2010)

[5] Data and scripts are available at https://github.com/ERATOMMSD/QIsPreferences.

3. Arcuri, A., Briand, L.: A practical guide for using statistical tests to assess randomized algorithms in software engineering. In: Proceedings of the 33rd International Conference on Software Engineering. pp. 1–10. ICSE '11, ACM, New York, NY, USA (2011)
4. Barros, M., Neto, A.: Threats to validity in search-based software engineering empirical studies. RelaTe-DIA **5** (01 2011)
5. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. Trans. Evol. Comp **6**(2), 182–197 (Apr 2002)
6. Durillo, J.J., Nebro, A.J.: jMetal: A Java framework for multi-objective optimization. Adv. Eng. Softw. **42**(10), 760–771 (Oct 2011)
7. Durillo, J.J., Nebro, A.J., Luna, F., Alba, E.: Solving three-objective optimization problems using a new hybrid cellular genetic algorithm. In: Parallel Problem Solving from Nature – PPSN X. pp. 661–670. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
8. Knowles, J.D., Corne, D.W.: Approximating the nondominated front using the pareto archived evolution strategy. Evol. Comput. **8**(2), 149–172 (Jun 2000)
9. Li, M., Chen, T., Yao, X.: A critical review of: "a practical guide to select quality indicators for assessing Pareto-based search algorithms in search-based software engineering": Essay on quality indicator selection for SBSE. In: Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results. pp. 17–20. ICSE-NIER '18, ACM, New York, NY, USA (2018)
10. Li, M., Yao, X.: Quality evaluation of solution sets in multiobjective optimisation: A survey. ACM Computing Surveys (12 2019)
11. Nebro, A.J., Durillo, J.J., Garcia-Nieto, J., Coello Coello, C.A., Luna, F., Alba, E.: SMPSO: A new PSO-based metaheuristic for multi-objective optimization. In: 2009 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making(MCDM). pp. 66–73 (March 2009)
12. Nebro, A.J., Durillo, J.J., Luna, F., Dorronsoro, B., Alba, E.: MOCell: A cellular genetic algorithm for multiobjective optimization. Int. J. Intell. Syst. **24**(7), 726–746 (Jul 2009)
13. Ravber, M., Mernik, M., Črepinšek, M.: The impact of quality indicators on the rating of multi-objective evolutionary algorithms. Appl. Soft Comput. **55**(C), 265–275 (Jun 2017)
14. Sayyad, A.S., Ammar, H.: Pareto-optimal search-based software engineering (posbse): A literature survey. In: 2013 2nd International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE). pp. 21–27 (May 2013)
15. Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures. Chapman & Hall/CRC, 5 edn. (2011)
16. Veček, N., Mernik, M., Črepinšek, M.: A chess rating system for evolutionary algorithms: A new method for the comparison and ranking of evolutionary algorithms. Information Sciences **277**, 656–679 (2014)
17. Wang, S., Ali, S., Gotlieb, A.: Cost-effective test suite minimization in product lines using search techniques. J. Syst. Softw. **103**(C), 370–391 (May 2015)
18. Wang, S., Ali, S., Yue, T., Li, Y., Liaaen, M.: A practical guide to select quality indicators for assessing Pareto-based search algorithms in search-based software engineering. In: Proceedings of the 38th International Conference on Software Engineering. pp. 631–642. ICSE '16, ACM, New York, NY, USA (2016)
19. Wohlin, C., Runeson, P., Hst, M., Ohlsson, M.C., Regnell, B., Wessln, A.: Experimentation in Software Engineering. Springer Publishing Company, Incorporated (2012)
20. Zitzler, E., Laumanns, M., Thiele, L.: SPEA2: Improving the strength Pareto evolutionary algorithm for multiobjective optimization. In: Evolutionary Methods for Design Optimization and Control with Applications to Industrial Problems. pp. 95–100. Athens, Greece (2001)