Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

Diagnostic Assessment of Search Controls and Failure Modes in Many-Objective Evolutionary Optimization

David Hadka dmh309@psu.edu
Department of Computer Science and Engineering, The Pennsylvania State University, University Park, 16802, USA

Patrick Reed pmr11@engr.psu.edu
Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park, 16802, USA

Abstract
The growing popularity of multiobjective evolutionary algorithms (MOEAs) for solving many-objective problems warrants the careful investigation of their search controls and failure modes. This study contributes a new diagnostic assessment framework for rigorously evaluating the effectiveness, reliability, efficiency and controllability of MOEAs as well as identifying their search controls and failure modes. The framework is demonstrated using the recently introduced Borg MOEA, ϵ-NSGA-II, ϵ-MOEA, IBEA, OMOPSO, GDE3, MOEA/D, SPEA2 and NSGA-II on 33 instances of 18 test problems from the DTLZ, WFG and CEC 2009 test suites. The diagnostic framework exploits Sobol's variance decomposition to provide guidance on the algorithms' non-separable, multi-parameter controls when performing many-objective search. This study represents one of the most comprehensive empirical assessments of MOEAs ever completed.

Keywords
Evolutionary computation, multiobjective optimization, many-objective optimization, search control, parameterization.

1 Introduction

Multiobjective evolutionary algorithms (MOEAs) have found favor by researchers and practitioners because of their ability to generate a Pareto approximation set in a single run for multiobjective optimization problems (MOPs). An increasingly large body of problems from diverse fields such as industrial, electrical, computer, civil and environmental engineering; aeronautics; finance; chemistry; medicine; physics and computer science are successfully employing MOEAs (Coello Coello et al., 2007). While in the majority of these domains MOEAs have been used predominately to solve two or three objective problems, there are growing demands for addressing higher dimensional problems yielding a growing research community in *many-objective optimization* (Fleming et al., 2005; Adra and Fleming, 2009).

Many-objective optimization involves the simultaneous optimization of four or more objectives. Several researchers have examined how problem difficulty is impacted by adding additional objectives to a problem. Fleming et al. (2005) studied the interaction between objectives. A *conflicting* interaction implies an improvement in one objective deteriorates another objective; on the other hand, an improvement to one objective

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

that improves another is *harmonious*. Teytaud (2006, 2007) proved the "true dimension" of a problem is the number of conflicting objectives. Brockhoff et al. (2007) observed that increasing the number of conflicting objectives in a problem changes the dominance relationships between objective vectors, resulting in either comparable objective vectors becoming incomparable or indifferent objective vectors becoming comparable. Furthermore, depending on whether the additional objective adds or removes deception, the expected runtime can either increase or decrease.

Additional challenges in many-objective optimization are related to the Pareto dominance relation. An objective vector *dominates* another if it is not worse in any objective and better in at least one. Two objective vectors that do not dominate one another are called *non-dominated* or *incomparable*. Farina and Amato (2004) and Fleming et al. (2005) observe that the proportion of locally non-dominated objective vectors tends to become large as the number of objectives increases, which is termed *dominance resistance* (Purshouse and Fleming, 2007). Dominance resistance generally increases with problem dimension and leads to a decline in selection pressure, thus slowing or stalling search. This observation led to several studies that redefine the dominance relation to provide more stringent dominance criteria. The *preferability* relation by Fonseca and Fleming (1998) is one of the first such contributions. Latter contributions include the relation *preferred* (Drechsler et al., 2001)[1], ϵ-*preferred* (Sülflow et al., 2007), k-*optimality* and its fuzzy counterpart (Farina and Amato, 2004), and *preference order* ranking (di Pierro et al., 2007). Classical methods of ranking non-dominated objective vectors, such as *average ranking*, have also been shown to provide competitive results (Corne and Knowles, 2007).

Teytaud (2006, 2007) shows that for large numbers of objectives (i.e., ≥ 10), the rule for selecting candidate solutions may not be more effective than random search. However, some modern MOEAs have the potential for significant search failures on problems with as few as four objectives (Ishibuchi et al., 2008a). At present, it is largely unknown if the dominance relations independently or jointly with search operators control failure modes for many-objective optimization. Also, the multivariate impacts of these algorithm choices implicitly bias search toward solutions which may or may not coincide with the decision maker's preferences. This search bias, for example, may improve convergence towards the compromise (knee) region(s), but result in undersampling of the extremities of the Pareto front. Since *a posteriori* methods require widespread search to provide the decision maker with a sufficient view of the Pareto surface, such biasing may be undesired and/or inappropriate.

Recent work on assessing search controls and ameliorating search failures has strongly focused on the interplay between diversity operators and selection pressure. Ishibuchi et al. (2008b) observed that the reduction in convergence to the Pareto front caused by dominance resistance increases the impact of diversity maintenance on population dynamics, a phenomenon termed *active diversity maintenance* (Purshouse and Fleming, 2007). Adra and Fleming (2009) suggest enabling or disabling the diversity maintenance operator in variation and survival selection in order to maintain an ideal spread factor. The spread factor identifies populations lacking diversity and those with excessive dispersal caused by divergence of solutions away from the Pareto front. With an ideal spread, rather than breaking ties with the crowding distance, which can hinder convergence, the variational operators and survival selection breaks ties randomly. Second, crowding in particular has known issues (Ishibuchi et al., 2008b), leading Kukkonen and Deb (2006) to introduce *pruning*, which recomputes crowding distances after

[1] The authors renamed this relation from *favour* to *preferred* in Sülflow et al. (2007)

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

each objective vector is removed during survival selection. Finally, Laumanns et al. (2002) proposed ϵ-*dominance* as a way to simultaneously preserve the proximity and diversity properties of MOEAs. ϵ-dominance redefines the dominance relation in order to encourage diversity by restricting the number of objective vectors permitted in a region of objective space. The success of several state-of-the-art MOEAs has been attributed to the use of ϵ-dominance (Deb et al., 2003; Sierra and Coello Coello, 2005; Kollat and Reed, 2006; Hadka and Reed, 2011).

To date, the complex dynamics of MOEAs when solving many-objective optimization problems has limited the analytical assessment of their strengths and weaknesses. Alternatively, with the advent of the DTLZ (Deb et al., 2001), WFG (Huband et al., 2006) and CEC 2009 (Zhang et al., 2009b) test problem suites, the systematic study of objective scaling through numerical experimentation has provided important insights into MOEA scalability for increasing objective dimensions. Khare et al. (2003) published the first study examining the effects of objective scaling on proximity and diversity using four DTLZ problems. Several additional experimental studies have been published using fixed or tuned parameters, as shown in Table 1. Purshouse and Fleming (2003, 2007) published the first study constructing control maps across a range of problem dimensions for the recombination and mutation operators for the Non-dominated Sorting Genetic Algorithm II (NSGA-II) (Deb et al., 2000) by sampling points on a grid from parameter space. They demonstrated that the parameterization sweet-spot migrates as the number of objectives increases. This result suggests that default parameterizations commonly used in the literature are not applicable to problems of varying dimensions.

More generally, Goh and Tan (2009) discuss the challenges in designing frameworks for the empirical analysis and performance assessment of MOEAs. They assert three important design requirements for any diagnostic framework: (1) multiple performance metrics covering the functional objectives of multiobjective optimization; (2) an adequate sample of problems; and (3) the ability to uncover pertinent parameter controls and dynamic search behavior within the algorithm. This study introduces a systematic framework for diagnosing the search capabilities of MOEAs while providing guidance on how the key multivariate interactions between an algorithm's parameters and its operators change as the number of objectives increases. This study represents one of the largest and most comprehensive computational experiments ever performed on MOEAs. Millions of algorithm runs using trillions of fitness function evaluations were executed to explore the design-space of state-of-the-art MOEAs. Such extensive experimentation supports the comparison of algorithms' best achieved metric values, their probabilities of attaining high-quality approximation sets, efficiency and controllability without biasing results to "tuned" rules for parameterization. Failures in this study for the first time imply failures in the MOEA's design — selection, variation operators, ranking, diversity maintenance, archiving, etc. and their interactions — rather than the synoptic analysis of poor parameterization effects, which has been the dominant focus of prior literature.

The remainder of this paper is organized as follows. The MOEAs and MOPs analyzed in this paper are introduced in Section 2 and Section 3, respectively. The diagnostic framework and our proposed measure of controllability are described in detail in Section 4. The results of a comparative analysis using this diagnostic framework is presented in Section 5 along with an analysis of search controls and failure modes. This paper concludes in Section 6 with a discussion of the impact of this work.

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

Table 1: List of prior comparison studies analyzing objective scaling for MOEAs. † marks algorithms modified specifically for handling many-objective optimization.

Algorithms	Problems	Objectives	Parameters	Reference
NSGA-II, SPEA2, PESA	DTLZ 1-3, 6	2-8	Tuned	Khare et al. (2003)
NSGA-II, MSOPS, RSO	Custom	2, 4, 6	Fixed	Hughes (2005)
NSGA-II, POGA†	DTLZ 1-4, 6-8	4-8	Tuned	di Pierro (2006)
NSGA-II, SPEA2, IBEA	DTLZ 1-7, WFG	2-4	Fixed	Wagner et al. (2007)
NSGA-II	DTLZ 1-3, 6	4-8	Tuned	Praditwong and Yao (2007)
NSGA-II, SPEA2, ϵ-MOEA	DTLZ 1-2	3-6	Fixed	Wagner et al. (2007)
NSGA-II, POGA	DTLZ 1-7	4-8	Tuned	di Pierro et al. (2007)
NSGA-II	DTLZ 2	3, 6, 12	Grid	Purshouse and Fleming (2003, 2007)
PESA-II	NK Landscapes	2, 5, 10	Fixed	Knowles and Corne (2007)
NSGA-II†	Knapsack	2, 4, 6, 8	Fixed	Ishibuchi et al. (2008a)
NSGA-II†	DTLZ2	6, 8, 12	Fixed	Adra and Fleming (2009)

Table 2: The MOEAs tested in this study.

Algorithm	Class	Reference
Borg MOEA	Adaptive multi-operator	(Hadka and Reed, 2011, This Issue)
ϵ-NSGA-II	Pareto front approximation	(Kollat and Reed, 2006)
ϵ-MOEA	Pareto front approximation	(Deb et al., 2002)
IBEA	Indicator-based	(Zitzler and Künzli, 2004)
OMOPSO	Particle swarm optimization	Sierra and Coello Coello (2005)
GDE3	Differential evolution	Kukkonen and Lampinen (2005)
MOEA/D	Aggregate functions	(Zhang et al., 2009a)
SPEA2	Baseline	(Zitzler et al., 2002a)
NSGA-II	Baseline	(Deb et al., 2000)

2 Multiobjective Evolutionary Algorithms

Evolutionary algorithms (EAs) are a class of search and optimization algorithms inspired by processes of natural evolution (Holland, 1975). Their ability to maintain a population of diverse solutions makes EAs a natural choice for solving multiobjective problems. The first MOEA to search for multiple Pareto-optimal solutions, the Vector Evaluated Genetic Algorithm (VEGA) was introduced in Schaffer (1984). VEGA was found to have problems similar to aggregation-based approaches, such as an inability to generate concave regions of the Pareto front. In 1989, Goldberg (1989a) suggested the use of Pareto-based selection, but this concept was not applied until 1993 in the Multiobjective Genetic Algorithm (MOGA) by Fonseca and Fleming (1993).

In the following years, several popular MOEAs with Pareto-based selection were published, including the Niched-Pareto Genetic Algorithm (NPGA) (Horn and Nafpliotis, 1993) and the Non-dominated Sorting Genetic Algorithm (NSGA) (Srinivas and Deb, 1993). Between 1993 and 2003, several *first-generation* MOEAs were introduced demonstrating important design concepts such as elitism, diversity maintenance and external archiving. Notable first-generation algorithms in addition to those already mentioned include the Strength Pareto Evolutionary Algorithm (SPEA) (Zitzler and Thiele, 1999), the Pareto-Envelope based Selection Algorithm (PESA) (Corne and Knowles, 2000) and the Pareto Archived Evolution Strategy (PAES) (Knowles and Corne, 1999). Most of these MOEAs have since been revised to incorporate more efficient algorithms and improved design concepts. For a more comprehensive overview of the historical development of MOEAs, please refer to the text by Coello Coello et al. (2007).

Since 2003, a number of MOEAs have been proposed in the literature for addressing more challenging problems and shortcomings in early MOEAs. Table 2 provides an overview of the MOEAs tested in this study, each of which are succinctly described

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

below. Readers seeking more details on the algorithms should reference their original cited works.. See (Hadka and Reed, 2011, This Issue) for more details on the Borg MOEA, which is the newest of the tested algorithms.

NSGA-II and SPEA2 The popular NSGA-II (Deb et al., 2000) and SPEA2 (Zitzler et al., 2002a) are two of the oldest MOEAs still in active use today (Coello Coello et al., 2007). Given their sustained popularity in the literature, they are included as baseline algorithms from which to compare more recent contributions.

ϵ-MOEA Issues like deterioration arise when finite population sizes force an MOEA to remove Pareto non-dominated solutions (Laumanns et al., 2002). As the proportion of Pareto non-dominated solutions increases as the number of objectives increases, the negative impacts of deterioration increases. Laumanns et al. (2002) introduced the ϵ-dominance relation as a way to eliminate deterioration by approximating the Pareto front, and also provided theoretical proofs of convergence and diversity for algorithms using this relation. ϵ-MOEA (Deb et al., 2002) is a popular steady-state MOEA that exploits the benefits of an ϵ-dominance archive. Note also that the Borg MOEA draws on ϵ-MOEA's highly efficient algorithmic structure in its implementation.

ϵ-NSGA-II ϵ-NSGA-II (Kollat and Reed, 2006) is another popular MOEA that combines NSGA-II, an ϵ-dominance archive, adaptive population sizing and time continuation (Goldberg, 1989b; Srivastava, 2002). ϵ-NSGA-II has been applied successfully to a broad array of real-world many-objective problems (Kollat and Reed, 2006, 2007; Kasprzyk et al., 2009; Ferringer et al., 2009; Kasprzyk et al., 2011; Kollat et al., 2011). In addition, many of its components influenced the design of the Borg MOEA (Hadka and Reed, 2011, This Issue).

Borg MOEA Given the variety of fitness landscapes and the complexity of search population dynamics, Vrugt and Robinson (2007); Vrugt et al. (2009) proposed adapting the operators used during multiobjective search based on their success in guiding search. Building on this work, the Borg MOEA (Hadka and Reed, 2011, This Issue) is designed for handling many-objective, multimodal problems using an auto-adaptive multioperator recombination operator to enhance search in a wide assortment of problem domains. This adaptive configuration of simulated binary crossover (SBX), differential evolution (DE), parent-centric recombination (PCX), unimodal normal distribution crossover (UNDX), simplex crossover (SPX), polynomial mutation (PM) and uniform mutation (UM) permits the Borg MOEA to quickly and reliably adapt to the problem's local characteristics and adjust as required throughout its execution (Hadka and Reed, 2011, This Issue). The Borg MOEA also introduces a variant of time continuation that combines ϵ-progress (i.e., the requirement that the approximation set is translated a minimum distance each generation) and adaptive population sizing to avoid search stagnation and maintain sufficient diversity to sustain search. The auto-adaptive multioperator recombination, adaptive population sizing, and time continuation components all exploit dynamic feedbacks from an ϵ-dominance archive to guarantee convergence and diversity throughout search, as shown by the theoretical analysis of Laumanns et al. (2002).

MOEA/D Using aggregation functions to convert a multiobjective problem into a single-objective problem has remained a popular search approach, but special care must be taken when designing the aggregation function to avoid widely acknowledged pitfalls. MOEA/D (Zhang et al., 2009a) is a recently-introduced MOEA that uses aggregate functions, but attempts to avoid the pitfalls in prior aggregation approaches

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

(Coello Coello et al., 2007; Wagner et al., 2007) by simultaneously solving many single-objective Chebyshev decompositions of many-objective problems in a single run. Since its introduction, MOEA/D has established itself as a benchmark for new MOEAs by winning the 2009 IEEE Congress on Evolutionary Computation (CEC 2009) competition (Zhang and Suganthan, 2009).

IBEA Indicator-based methods replace the Pareto dominance relation with an indicator function intended to guide search towards regions of interest (Ishibuchi et al., 2010). IBEA uses the hypervolume measure, which avoids active diversity maintenance by not using an explicit diversity-preserving mechanism. Diversity is instead promoted through the hypervolume measure itself (Wagner et al., 2007). One potential downfall to hypervolume-based methods is the computational complexity of calculating the hypervolume measure on high-dimensional problems, although it should be noted that Ishibuchi et al. (2010) have proposed an approximation method to reduce the computational complexity.

GDE3 GDE3 (Kukkonen and Lampinen, 2005) is a multiobjective variant of differential evolution (DE). GDE3 (and DE in general) is notable for *rotationally invariant* operators — they produce offspring independent of the orientation of the fitness landscape — which is important for problems with high degrees of conditional dependence among its decision variables (Iorio and Li, 2008). GDE3 was a strong competitor in the CEC 2009 competition (Zhang and Suganthan, 2009).

OMOPSO OMOPSO (Sierra and Coello Coello, 2005) is one of the most successful multiobjective particle swarm optimization (PSO) algorithms to date. It is notable for being the first multiobjective PSO algorithm to include ϵ-dominance as a means to solve many-objective problems. OMOPSO thus provides a representative baseline from the PSO class of algorithms.

3 Multiobjective Test Problems

The 33 instances of 18 unconstrained, real-valued multiobjective test problems listed in Table 3 are used in this study. Also shown are the ϵ values used for ϵ-dominance. The UF1-UF13 problems are the unconstrained problems used during the IEEE Congress on Evolutionary Computation (CEC) competition held in 2009 (Zhang et al., 2009b). UF11 and UF12 are rotated instances of the 5D DTLZ2 and DTLZ3 test problems, respectively (Deb et al., 2001). UF13 is the 5D WFG1 test problem (Huband et al., 2006). The DTLZ problems are from a set of scalable test problems. In this study, these problems are tested with 2, 4, 6 and 8 objectives. Table 3 also lists the ϵ values used for ϵ-dominance. For the scalable DTLZ test problems, the ϵ values used were 0.01, 0.15, 0.25 and 0.35 for 2, 4, 6 and 8 objectives, respectively.

Restricting the scope of this study to a maximum of 8 objectives and unconstrained problems was necessary given the computational requirements of performing such a large scale study. In addition, some algorithms (MOEA/D) are not readily applicable to standard constraint-handling techniques. Nevertheless, our future work will apply the diagnostic framework presented in Section 4 to other algorithms and problem domains, such as constrained problems, dynamic landscapes, noisy landscapes, and goal/preference guided MOEAs.

The conference version of the DTLZ suite (Deb et al., 2002) omits two problems and relabels another, this study along with most others use the problems and names defined in Deb et al. (2001). DTLZ5 and DTLZ6 were omitted since the original problem

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

Table 3: The problems used in the comparative study along with key properties.

Problem	M	L	Properties	ϵ
UF1	2	30	Complicated Pareto Set	0.001
UF2	2	30	Complicated Pareto Set	0.005
UF3	2	30	Complicated Pareto Set	0.0008
UF4	2	30	Complicated Pareto Set	0.005
UF5	2	30	Complicated Pareto Set, Discontinuous	0.000001
UF6	2	30	Complicated Pareto Set, Discontinuous	0.000001
UF7	2	30	Complicated Pareto Set	0.005
UF8	3	30	Complicated Pareto Set	0.0045
UF9	3	30	Complicated Pareto Set, Discontinuous	0.008
UF10	3	30	Complicated Pareto Set	0.001
UF11	5	30	DTLZ2 5D Rotated	0.2
UF12	5	30	DTLZ3 5D Rotated	0.2
UF13	5	30	WFG1 5D	0.2
DTLZ1	2-8	M+4	Multimodal, Separable	0.01-0.35
DTLZ2	2-8	M+9	Concave, Separable	0.01-0.35
DTLZ3	2-8	M+9	Multimodal, Concave, Separable	0.01-0.35
DTLZ4	2-8	M+9	Concave, Separable	0.01-0.35
DTLZ7	2-8	M+19	Discontinuous, Separable	0.01-0.35

definitions produce Pareto fronts differing from the published analytical solutions with four or more objectives. This issue was identified by Huband et al. (2006) and corrected in Deb and Saxena (2006) by including additional problem constraints. DTLZ8 and DTLZ9 also include side constraints and were consequently omitted from this study.

4 Diagnostic Framework

The primary contribution of this study is a diagnostic framework for robustly comparing how MOEA operators, their parameterization, and the interactions between these factors influence their successes and failures in many-objective optimization. Section 4.1 briefly discusses common performance metrics used in prior studies. Section 4.2 defines the best attained approximation set, effectiveness and controllability metrics used by this diagnostic framework. Section 4.3 introduces variance decomposition of controls for analyzing the multivariate interactions between parameters. Section 4.4 outlines the computational experiment performed in this study. Table 4 identifies common notations used throughout this section.

4.1 Performance Metrics

The present MOEA literature does not show consensus on how multiobjective search performance should be evaluated and/or compared. Zitzler et al. (2002b,c) contend that the number of unary performance metrics required to determine if one approximation set is preferred over another must be at least the number of objectives in the problem[2]. Because different MOEAs tend to perform better in different metrics (Bosman and Thierens, 2003), Deb and Jain (2002) suggest only using metrics for the two main

[2]Binary indicators alleviate this and other issues, but are more difficult to handle and would hinder comparability between studies.

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

Table 4: Notation used in study.

Symbol	Description
M	Number of objectives
$s \in S$	A seed from the set of seeds
$p \in \mathcal{P}$	A parameter set in the parameter block
A	One of the studied algorithms
A_p^s	A single run of the algorithm A using parameter set p and seed s; returns an approximation set
A_p	Shorter version of A_p^s implying a single seed s is used
$\mathcal{M}(A_p^s)$	Performance metric applied to the approximation set from a single run
$\mathcal{M}^*$	Target metric value (i.e., best achievable metric value given a reference set)

functional objectives of MOEAs: proximity and diversity. Knowles and Corne (2002) suggest the *hypervolume* metric because it is compatible with the outperformance relations, scale independent, intuitive, and can reflect the degree of outperformance between two approximation sets.

In this study, we acknowledge the challenge of performance assessment and have computed a broad range of performance metrics, including hypervolume, generational distance (GD), inverse generational distance (IGD), additive epsilon indicator (ϵ_+-indicator) and spread (Coello Coello et al., 2007; Deb et al., 2000). We have chosen to present results only for the GD, hypervolume, and ϵ_+-indicator metrics for the following reasons. GD is the average distance from objective vectors in the approximation set to the nearest objective vector in the reference set, thus representing proximity. Hypervolume measures the volume of objective space covered/dominated by the approximation set, thus representing a combination of proximity and diversity. ϵ_+-indicator is the smallest distance the approximation set must be translated so that every objective vector in the reference set is covered. This identifies situations in which the approximation set contains one or more outlying objective vectors with poor proximity. If an approximation set for the most part shows good proximity except for a few objective vectors, then that approximation set is not *consistent*. This study therefore defines the three main functional objectives of MOEAs as proximity, diversity and consistency. Figure 1 provides a graphical representation of the importance of the ϵ_+-indicator and consistency.

In order to handle performance metrics based on maximization and minimization, all performance metrics are normalized. This normalization converts all performance metrics to reside in the range $[0, 1]$ with 1 representing the optimal value achievable using a generated reference set. Hypervolume is normalized by

$$\mathcal{M}(A_p^s) = \widehat{\mathcal{M}}(A_p^s)/\mathcal{M}^*, \tag{1}$$

where $\widehat{\mathcal{M}}$ represents the raw metric value. GD and the ϵ_+-indicator are normalized by

$$\mathcal{M}(A_p^s) = \max(1 - \widehat{\mathcal{M}}(A_p^s), 0). \tag{2}$$

Reference sets were generated using the analytical solutions to all test problems.

4.2 Search Control Metrics

Whereas the performance metrics discussed in Section 4.1 compare the quality of approximation sets from single runs, they are only applicable to fixed parameterizations. In this study we propose instead the following metrics for statistically sampled ensembles of approximation sets and their corresponding performance metrics to provide

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

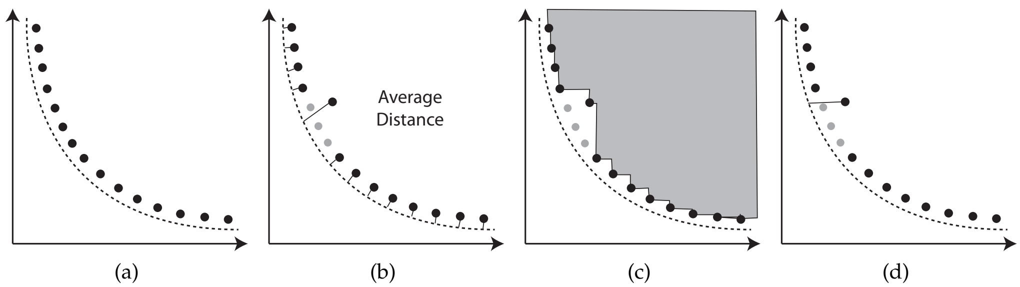


Figure 1: Demonstrates the importance of ϵ-indicator as a measure of consistency. (a) A good approximation set to the reference set, indicated by the dashed line. (b) Generational distance averages the distance between the approximation set and reference set, reducing the impact of large gaps. The missing points are shaded light gray. (c) The change in hypervolume due to a gap is small relative to the entire hypervolume. (d) ϵ-Indicator easily identifies the gap, reporting a metric 2-3 times worse in this case.

guidance on an MOEA's utility. Our diagnostic framework classifies an MOEA's utility using four measures: best achieved value, probability of attainment, efficiency and controllability.

Best Achieved Value The majority of studies report the best achieved end-of-run performance metric value. However, unlike the majority of studies where results are based on fixed or tuned parameters, our best attained result is drawn from a large statistical sampling of the full feasible parameterization ranges for all of the major operators in each algorithm in order to provide a rigorous measure of an MOEA's best performance.

$$\text{Best Achieved Value} = \max_{p \in \mathcal{P}} \mathcal{M}(A_p) \tag{3}$$

Probability of Attainment While the best achieved value is an absolute measure of an MOEA's search quality, the reliability of an algorithm is a stronger indicator of an MOEA's utility. This is particularly important on rotated, multi-modal, many-objective problems where an MOEA may be capable of producing quality end-of-run approximation sets, but the probability of doing so is low. We propose measuring an MOEA's reliability with the probability the end-of-run approximation set surpasses an attainment threshold, α. From the set of parameters $\mathcal{P}$, the set of parameters surpassing this attainment threshold is

$$\mathcal{P}^{\alpha} = \{p \in \mathcal{P} : \mathcal{M}(A_p) \geq \alpha\} . \tag{4}$$

From this, the probability of attainment is defined by

$$\text{Probability of Attainment} = \frac{|\mathcal{P}^{\alpha}|}{|\mathcal{P}|}. \tag{5}$$

Efficiency MOEAs that achieve high attainment probabilities with fewer objective function evaluations are preferred over those that require more time to search. Efficiency measures the minimum number of objective function evaluations (NFE) required to achieve a high probability of attainment. Given a range R of NFE values, we define a band of statistically sampled parameterizations within that range as

$$\mathcal{B}_R = \{p \in \mathcal{P} : \text{NFE}(p) \in R\} , \tag{6}$$

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

and the subset of parameterizations in that band surpassing the attainment threshold as

$$\mathcal{B}_R^\alpha = \{p \in \mathcal{B}_R : \mathcal{M}(A_p) \geq \alpha\}\,. \tag{7}$$

Efficiency is defined as the minimum NFE band R such that 90% of the parameters in the band surpass the attainment threshold:

$$\text{Efficiency} = \min\left\{R : \frac{|\mathcal{B}_R^\alpha|}{|\mathcal{B}_R|} \geq 0.9\right\}, \tag{8}$$

where $R = [\Delta i, \Delta(i+1)]$ for $i = \{0, \ldots, 99\}$ and $\Delta = 10000$. Note the similarities between these equations and those for the probability of attainment. The choice of 90% is based on our efforts to maintain consistency and rigor across our performance measures. In the context of this specific study, there were no significant differences in efficiency if 50% and 75% thresholds were stipulated.

Controllability Lastly, we are interested in the distribution of the parameters in $\mathcal{P}^\alpha$. Controllable algorithms are those which exhibit *sweet spots*, or regions in parameter space with high attainment probabilities. The *correlation dimension* (Grassberger and Procaccia, 1983) of $\mathcal{P}^\alpha$ is our measure of controllability. Hence, controllability is computed by

$$\text{Controllability} = \lim_{r \to 0} \frac{\ln(C(r))}{\ln(r)}, \tag{9}$$

where

$$C(r) = \frac{1}{N(N-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} H(r - |p_i - p_j|) \tag{10}$$

with $p_i, p_j \in \mathcal{P}^\alpha$, $N = |\mathcal{P}^\alpha|$ and H is the Heaviside function defined by

$$H(u) = \begin{cases} 0 & \text{if } u < 0 \\ 1 & \text{if } u \geq 0. \end{cases} \tag{11}$$

Conceptually, $C(r)$ is the average fraction of parameter sets within a radius r of each other. The growth of $C(r)$ with respect to r reflects dimensionality since higher dimensional spaces permit more opportunities for points to be close (Baker and Gollub, 1990). As shown in Figure 2, rather than computing (9) directly, it is recommended to instead compute the slope where the correlation dimension estimate $\ln(C(r))/\ln(r)$ is relatively constant (this region is called the *plateau region* in the literature) (Nayfeh and Balachandran, 1995).

To compute (9), the effective parameters $\mathcal{P}^\alpha$ are first normalized to reside within the unit hypercube. The $N(N+1)/2$ pairwise distances between effective parameters are computed and stored in an array. $C(r)$ from (10) is computed for various $r \in [0,1]$ by referencing distances in this stored array. Next, the plateau region is identified, as shown in Figure 2. Let $\mathcal{R} = \{r : r_{\text{min}} \leq r \leq r_{\text{max}}\}$ be the sampled values of r within some bounds. The linearity of $\ln(C(r))$ versus $\ln(r)$ is determined by computing the correlation coefficient (Edwards, 1993)

$$\rho = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}, \tag{12}$$

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

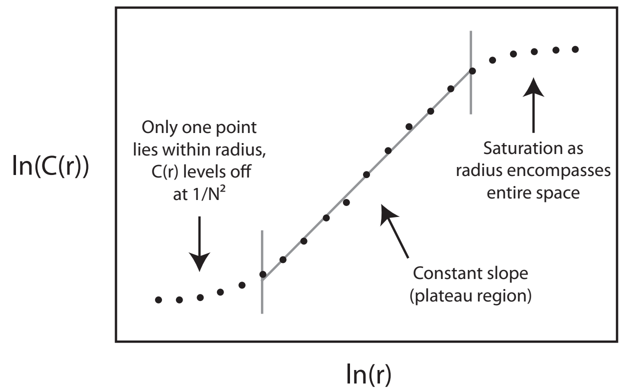


Figure 2: The correlation dimension is the slope where the correlation dimension estimate $\ln(C(r))/\ln(r)$ is relatively constant (this region is called the *plateau region* in the literature). As indicated, small and large radii do not reflect dimensionality.

where the summations are over the values $r \in \mathcal{R}$, $n = |\mathcal{R}|$, $x = \ln(r)$ and $y = \ln(C(r))$. Searching for the largest bounds, $r_{\max} - r_{\min}$, with $|\rho| \geq 1 - \xi$ identifies the plateau region. This study used $\xi = 0.001$ to ensure a high degree of linearity. Finally, the slope of the identified plateau region and the estimation of (9) is calculated using linear least squares regression (Edwards, 1993)

$$m = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}, \tag{13}$$

with the same variables as (12).

In summary, controllability measures the correlation between effective parameters. Thus, larger controllability values indicate increasingly larger perturbations to an effective parameter set will still result in good performance, which indicates the existence of sweet spots. The existence of sweet spots is necessary for the effective control of search via parameterization. Without sweet spots, adapting parameters becomes hard since effective parameters are like needles in a haystack — small perturbations to effective parameters will likely result in poor performance.

4.3 Variance Decomposition of Controls

The highly non-linear nature of MOEAs emerges from complex interactions between their operators and their parameterization, which has limited the analysis of generalized MOEA behavior. Most studies to date only examine one or two parameters in isolation (Harik and Lobo, 1999). However, recent advances in sensitivity analysis have introduced techniques for computing all parameter effects and their multivariate interactions more reliably and with fewer parametric assumptions relative to traditional methods like analysis of variance (ANOVA).

Variance decomposition attributes to each parameter the percentage it contributes to an output ensemble's variance. *First-order* effects represent variation caused solely by a single parameter. *Second-order* and *higher-order* interaction effects represent variation caused by two or more parameters in conjunction. *Total-order* effects represent for each parameter the summation of its first-order and all higher-order effects.

While ANOVA has been traditionally used to capture first- and second-order effects, the variance decomposition method developed by I.M. Sobol' with modifications

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

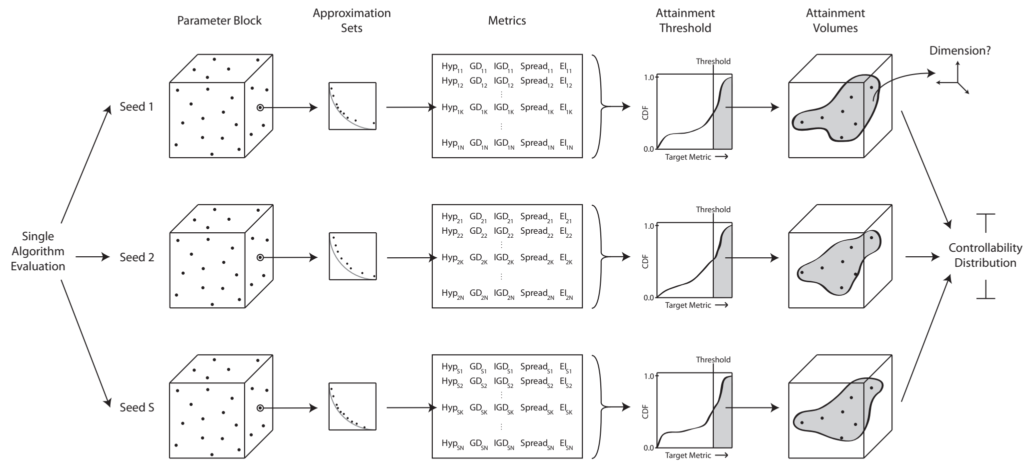


Figure 3: For each algorithm, a Sobol' sequence-based statistical sampling of its parameters is generated (i.e., the parameter block). Each parameter set in the parameter block is evaluated using multiple random number seed trials ($S = 50$) to improve the statistical quality of our results. From the resulting non-dominated approximation sets, the corresponding performance metrics are computed. An attainment threshold retains all parameter settings surpassing the threshold value, which are then used to compute the probability of attainment, efficiency and controllability measures.

by Saltelli et al. (2008) provides many advantages. First, using the implementation in Saltelli et al. (2008), the total-order effects can be computed with little additional cost over Sobol's original implementation. Second, whereas uniform random sampling of parameters yields a sampling error growth rate of $1/\sqrt{N}$, sampling parameters with Sobol's quasi-random sequence generator yields an error growth rate of $1/N$, a significant improvement in convergence (Tang et al., 2007). In this study, $N = |\mathcal{P}|$. Third, the rank-ordering of parameters by Sobol's method has been observed in practice to be more reliable and stable than ANOVA (Tang et al., 2007). Finally, Sobol's method is model independent and only assumes parameter independence. ANOVA, on the other hand, assumes normally-distributed model responses, homoscedasticity, and independence of cases.

For these reasons, Sobol's variance decomposition is used in this study to identify an MOEA's key parameters and investigate the multivariate interactions between its control parameters. Error estimates are determined using bootstrapping. A more detailed discussion of Sobol's variance decomposition and bootstrapping is provided in the appendix.

4.4 Computational Experiment

This study applies the nine algorithms listed in Table 2 to the 33 test problem instances listed in Table 3. Figure 3 depicts the overall outline of this computational experiment, which is described in detail below. To permit Sobol's variance decomposition for each algorithm, a *parameter block* consisting of $1000(2P + 2)$ parameter sets is generated using a Sobol' sequence-based statistical sampling method, where P is the number of parameters controlling the algorithm. For each parameter set in the parameter block, the algorithm is run 50 times using different initial pseudo-random number generator

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

seeds for each problem instance. The same parameter block is used across all seeds and problem instances for each algorithm. The result of each run is a Pareto approximation set which is evaluated using the performance metrics discussed in Section 4.1. The multiple random number seed trials render the results independent of the initial population and improve the statistical quality of our results.

After all the data is collected, the search control metrics and variance decomposition of controls are computed. Each parameter block is analyzed to identify only those runs surpassing an 75%-attainment threshold relative to the known reference sets. The resulting *attainment volume* is used to compute the probability of attainment, efficiency and controllability search control metrics. Along with the best achieved value, these measures of algorithmic utility can be used to make observations of the current state-of-the-field for solving many-objective problems. Additionally, our framework utilizes Sobol's variance decomposition to rigorously assess algorithms search controls while simultaneously providing insights into the multivariate interactions between parameters and operators. Our proposed use of variance decomposition clearly characterizes the effect of objective scaling on MOEA search.

The range of sampled parameter values is taken from Hadka and Reed (2011, This Issue). The number of fitness evaluations was sampled between $[10000, 1000000]$ in order to permit tractable execution times while providing meaningful results. The population size, offspring size, and archive sizes are all sampled between $[10, 1000]$. This range was chosen to encompass the commonly employed "rule-of-thumb" population sizes in MOEA parameterization recommendations. Mutation rate, crossover rate, and step size encompass their entire feasible ranges of $[0, 1]$. Distribution indices for SBX and PM range between $[0, 500]$, which is based on the sweet spot identified by Purshouse and Fleming (2007).

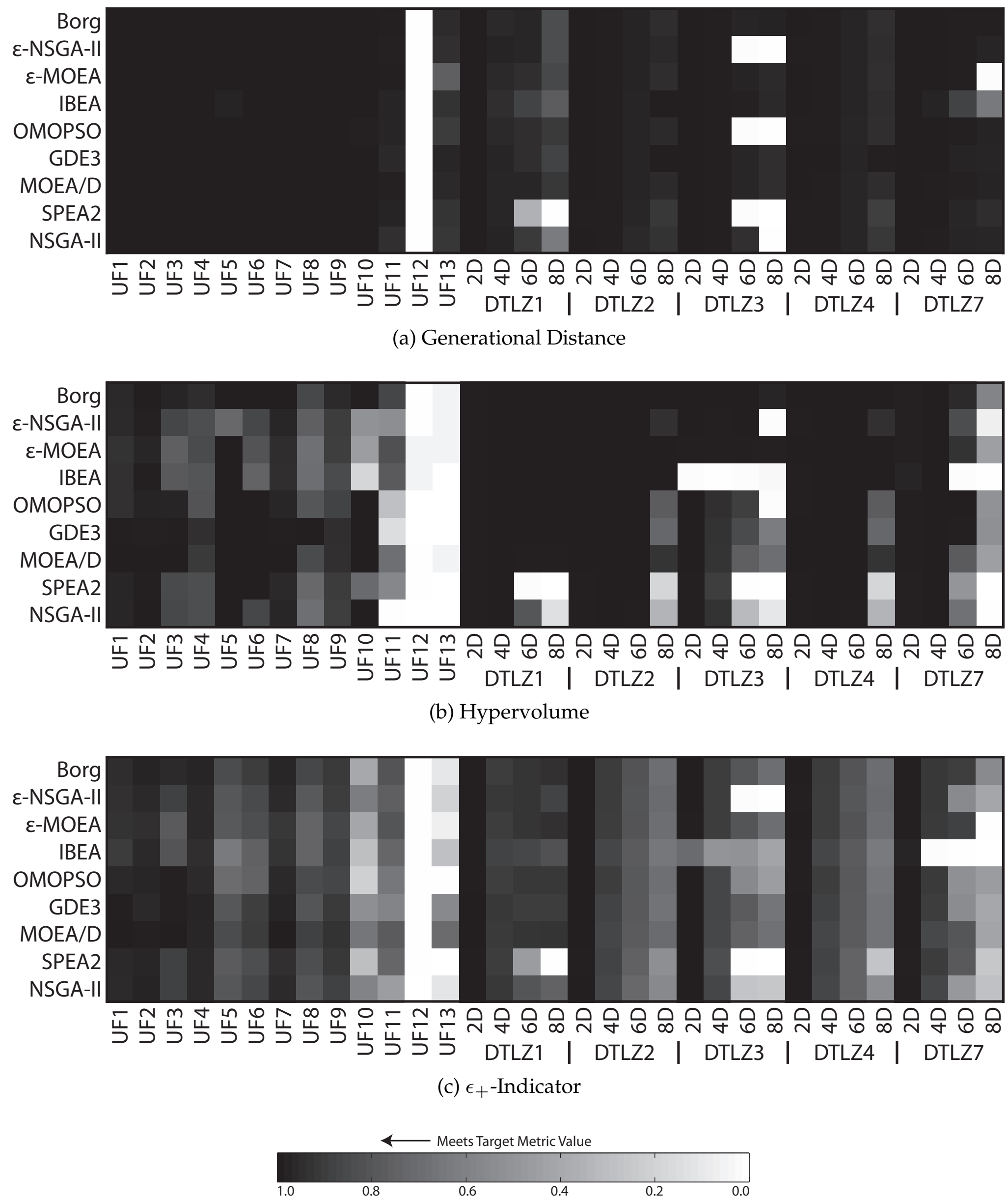
The above experiment was executed on the CyberStar computing cluster at the Pennsylvania State University, which consists of 512 2.7 GHz processors and 1536 2.66 GHz processors. In total, 280 million algorithm runs were executed requiring approximately 225 years of computational effort. To the best of our knowledge, this is the most extensive and comprehensive comparison study of MOEAs to date. Consequently, our results do not rely on fixed or tuned parameters and provides a state-of-the-field baseline for many-objective evolutionary optimization. While the computational expenditure for this study is high, it has freed our analysis and results from restrictive assumptions, and is the first robust analysis that statistically samples the design space of MOEAs.

5 Results and Discussion

Figs. 4, 5, 6 and 7 show the best achieved value, probability of attainment, efficiency and controllability measures, respectively, for the 33 test problem instances. Each plot contains three horizontal subplots containing the generational distance (GD), hypervolume, and ϵ_+-indicator performance metrics. Each subplot is composed of shaded squares corresponding to the problem (x-axis) and the algorithm (y-axis). The interpretation of the shading depends on the individual plot, but in all cases black represents the ideal result and white the worst result. All shadings are scaled linearly as indicated in the legends.

Figure 4 shows for each MOEA its overall best achieved metric value for the three performance metrics. Dark regions indicate at least one of the sampled parameter sets attained performance metric values very near to the target metric value. Starting with GD, which measures the average distance from objective vectors in the approximation

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053



(a) Generational Distance

(b) Hypervolume

(c) ϵ_+-Indicator

Figure 4: The overall best performance for each algorithm on each problem instance is illustrated as the percentage of target metric value achieved. The targets for each problem are based on their true reference sets. Black regions indicate there exists at least one parameter set that yielded near-optimal metric values. White regions indicate no such parameter set exists.

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

set to the nearest neighbor in the reference set, we observe that at least one parameter set was able to attain near-optimal convergence to the reference set for most problem instances. We observe that all of the algorithms had difficulty on the UF12 problem from the CEC 2009 test suite, and ϵ-NSGA-II, OMOPSO and SPEA2 had difficulty on the 6 and 8 dimension cases of the DTLZ3 problem. In addition, NSGA-II struggled on the 8D DTLZ3 instance and SPEA2 struggled on the 8D DTLZ1 instance. This indicates that apart from these few exceptions, the majority of the tested algorithms are capable of producing at least one approximation set in close proximity to the reference set. While GD measures proximity to the reference set, a non-diverse population covering only a small fraction of the reference set can receive near-optimal GD values. In other words, GD provides no information about diversity.

The hypervolume performance metric, which measures the volume of space dominated by the approximation set, combines proximity and diversity into a single evaluation metric. Again, the majority of the tested algorithms are able to generate at least one approximation set with a hypervolume near the reference set. First, we observe low hypervolume values on UF11, UF12 and UF13. Given the near-optimal GD values on UF11 and UF13, this indicates the MOEAs struggle to maintain a diverse set of solutions on these problem instances. This loss in diversity is also apparent for IBEA on DTLZ3 and DTLZ7. On DTLZ3, IBEA struggles to maintain a diverse approximation set regardless of problem dimension. This indicates a significant search failure for IBEA, particularly given the fact that IBEA is based on the hypervolume indicator. The Borg MOEA is able to achieve near-optimal hypervolume values for the majority of the tested problem instances, only struggling on UF12, UF13 and 8D DTLZ7. Borg's ability to maintain a diverse approximation set is aided by its use of an ϵ-dominance archive.

The last metric shown in Figure 4 is ϵ_+-indicator. The ϵ_+-indicator highlights the existence of gaps in the Pareto fronts (i.e., consistency as illustrated in Figure 1). The ϵ_+-indicator highlights the difficulty of UF12 and UF13 as detected by GD and hypervolume. A clear pattern emerges on the DTLZ problems showing a degradation in performance of the algorithms at higher problem dimensions. The Borg MOEA, ϵ-NSGA-II, ϵ-MOEA and MOEA/D show a slight advantage, particularly on higher-dimensional DTLZ problem instances.

Combining these three performance metrics provides a clear indication as to the quality of an approximation set. A favorable GD value implies good proximity, a favorable hypervolume implies good diversity with proximity, and a favorable ϵ_+-indicator value implies good consistency (i.e., the absence of poorly approximated regions in the approximation set). As an example, an MOEA exhibiting good GD but poor ϵ_+-indicator values implies some regions of the reference set are approximated poorly. Tradeoffs between the various algorithms with respect to the functional objectives of MOEAs is evident; however, the Borg MOEA shows the most successful results across all functional objectives. Alternatively, IBEA, SPEA2 and NSGA-II struggled to produce diverse approximation sets on many-objective problems.

Readers should note that in addition to the tested algorithms, random search was used to establish a baseline comparison. The random search baseline was established by randomly generating the same number of solutions as were evaluated by the MOEAs and adding them to an ϵ-dominance archive using the same ϵ values as Borg and OMOPSO. The performance metrics were computed for the approximation sets generated by random search. In all cases excluding UF12, where all algorithms failed, the MOEAs outperformed random search. This fact is important as it implies the MOEAs are performing non-trivial search.

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

It is interesting to note the difficulty observed on UF12. UF12 is the rotated version of the 5D DTLZ3 problem originally used in the CEC 2009 competition. This suggests that state-of-the-art MOEAs still show significant search failures on rotated multimodal many-objective problems. This highlights the need for further advancements in this area.

Many studies feature the best observed metric, but such cherry picking of parameters poorly reflects a user's ability to utilize an MOEA in real-world applications where search failures can have actual economic costs. Recall that this study uses an 75%-attainment threshold when calculating the probability of attainment. The probability of attainment, which is the percentage of sampled parameter sets that are able to achieve 75% of each problem instance's reference set, is shown in Figure 5. Black identifies cases where the majority of the parameter sets sampled are successful in attaining high quality approximation sets.

Starting with GD in Figure 5, we observe that all algorithms exhibit high attainment probabilities on most UF problems and all tested dimensions of DTLZ2, DTLZ4 and DTLZ7. For these cases, the majority of the parameters sampled produce results with a high level of proximity. However, this does not hold for DTLZ1 and DTLZ3. The majority of the tested MOEAs show low attainment probabilities, even on 2D and 4D DTLZ3. The Borg MOEA, ϵ-NSGA-II, ϵ-MOEA and NSGA-II were the only MOEAs that retained high attainment probabilities on 2D and 4D DTLZ3.

In addition, the hypervolume and ϵ_+-indicator values show diversity and consistency are issues. With the exceptions of UF1, UF2, UF4, UF7 and lower-dimensional DTLZ problem instances, the tested algorithms were not reliably capable of producing well-spread and consistent approximation sets. The Borg MOEA, ϵ-NSGA-II, ϵ-MOEA and MOEA/D provide better diversity and consistency than the other MOEAs, but even these struggle on higher-dimensional instances.

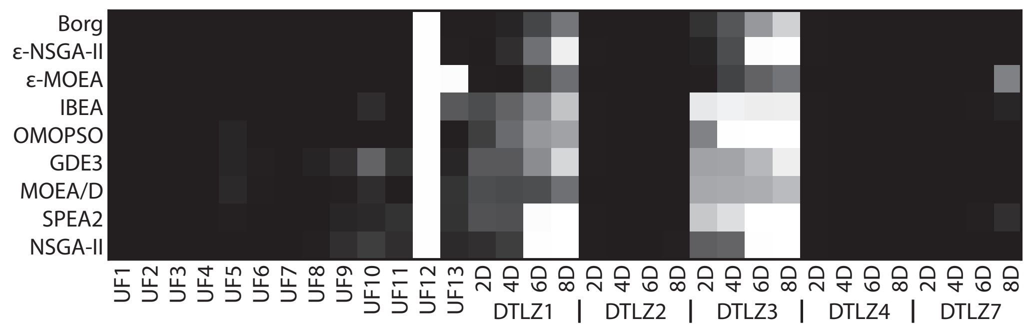
The general trend across all of the algorithms' low attainment probabilities on DTLZ1 and DTLZ3 suggests multimodal problems can cause significant search failure. In combination, Figure 4 and Figure 5 show that these algorithms can attain high quality solutions, but the probability of it occurring using commonly selected parameters decreases significantly as the objective space dimension increases.

Efficiency reflects the amount of effort expended by the MOEA, in terms of the number of objective function evaluations (NFE), to produce approximation sets surpassing the 75% attainment threshold. Figure 6 shows the efficiency results, where black regions indicate cases where the MOEA required fewer NFE and white indicates the MOEA failed to surpass the attainment threshold. Looking at GD, the majority of the tested MOEAs produced approximation sets with good proximity with 200k or fewer NFE. The few exceptions are NSGA-II, SPEA2, OMOPSO, IBEA and ϵ-NSGA-II on DTLZ3. NSGA-II, SPEA2 and ϵ-NSGA-II also struggled on higher-dimensional DTLZ1 in terms of efficiency. MOEA/D struggled on UF13 and 8D DTLZ7.

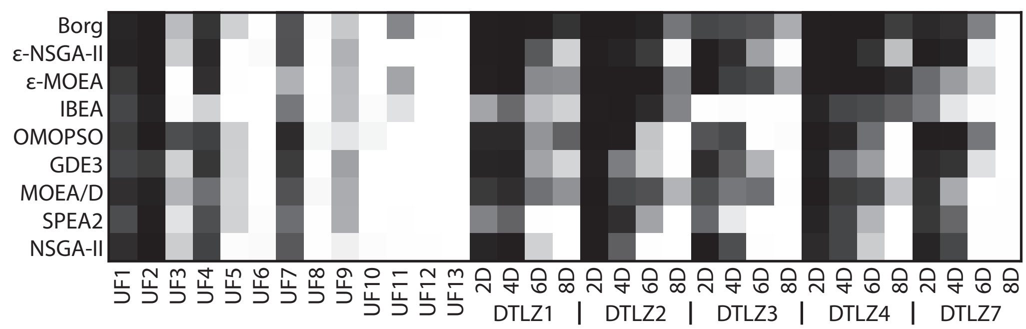
Looking at hypervolume and ϵ_+-indicator, low efficiencies occur on UF6, UF8, UF10-UF13 and higher-dimensional DTLZ problem instances. Comparing these results to Figure 5, reduced efficiency corresponds with low attainment probabilities. If the algorithm fails to reliably generate approximation sets surpassing the attainment threshold, they will also be marked with low efficiency. On the scalable DTLZ instances, we observe a rapid loss in efficiency as the problem dimension increases. The Borg MOEA, ϵ-MOEA and MOEA/D are the only MOEAs with high efficiency on the higher-dimensional multimodal DTLZ1 and DTLZ3 instances.

Although the reliability and efficiency of the algorithms are important, it is equally

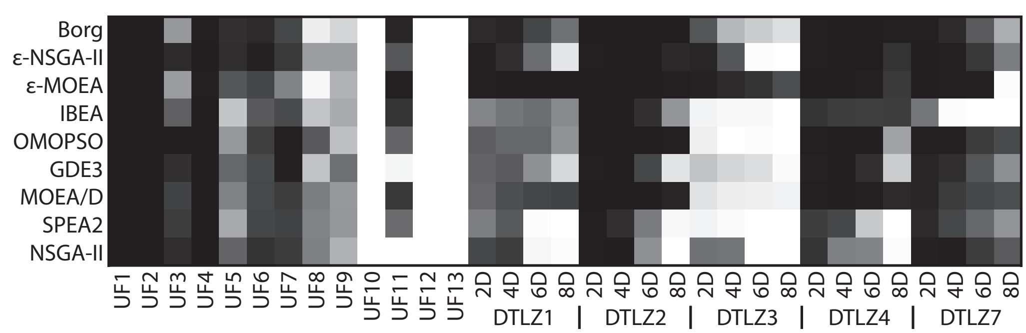
Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053



(a) Generational Distance



(b) Hypervolume



(c) ϵ_+-Indicator

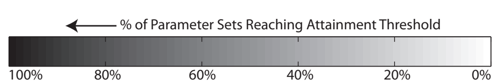


Figure 5: The probability of attainment results illustrate the percent of parameter sets for each algorithm that yielded end-of-run metric values surpassing a 75%-attainment threshold. Black regions indicate large success rates while white regions indicate low success rates.

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

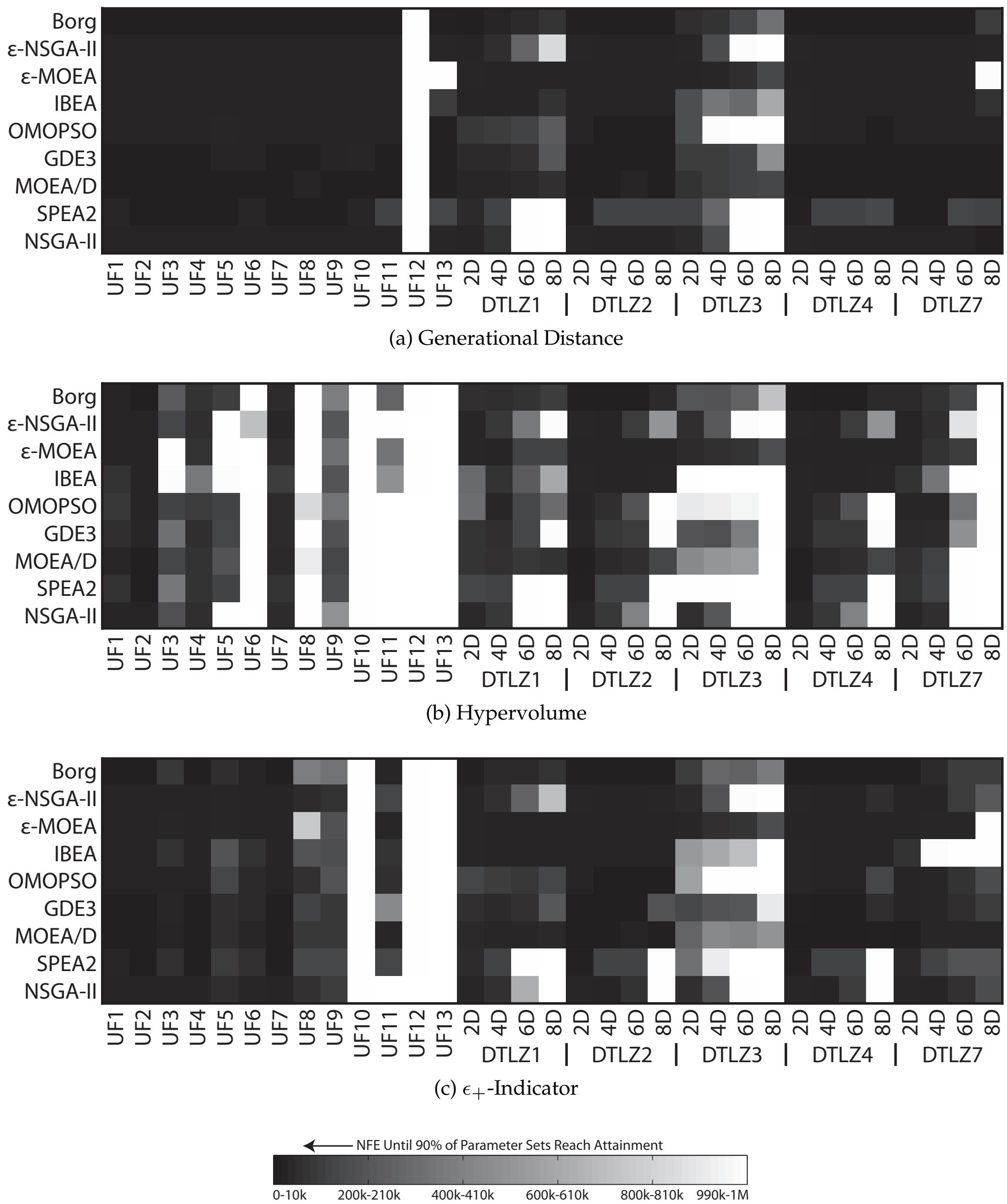
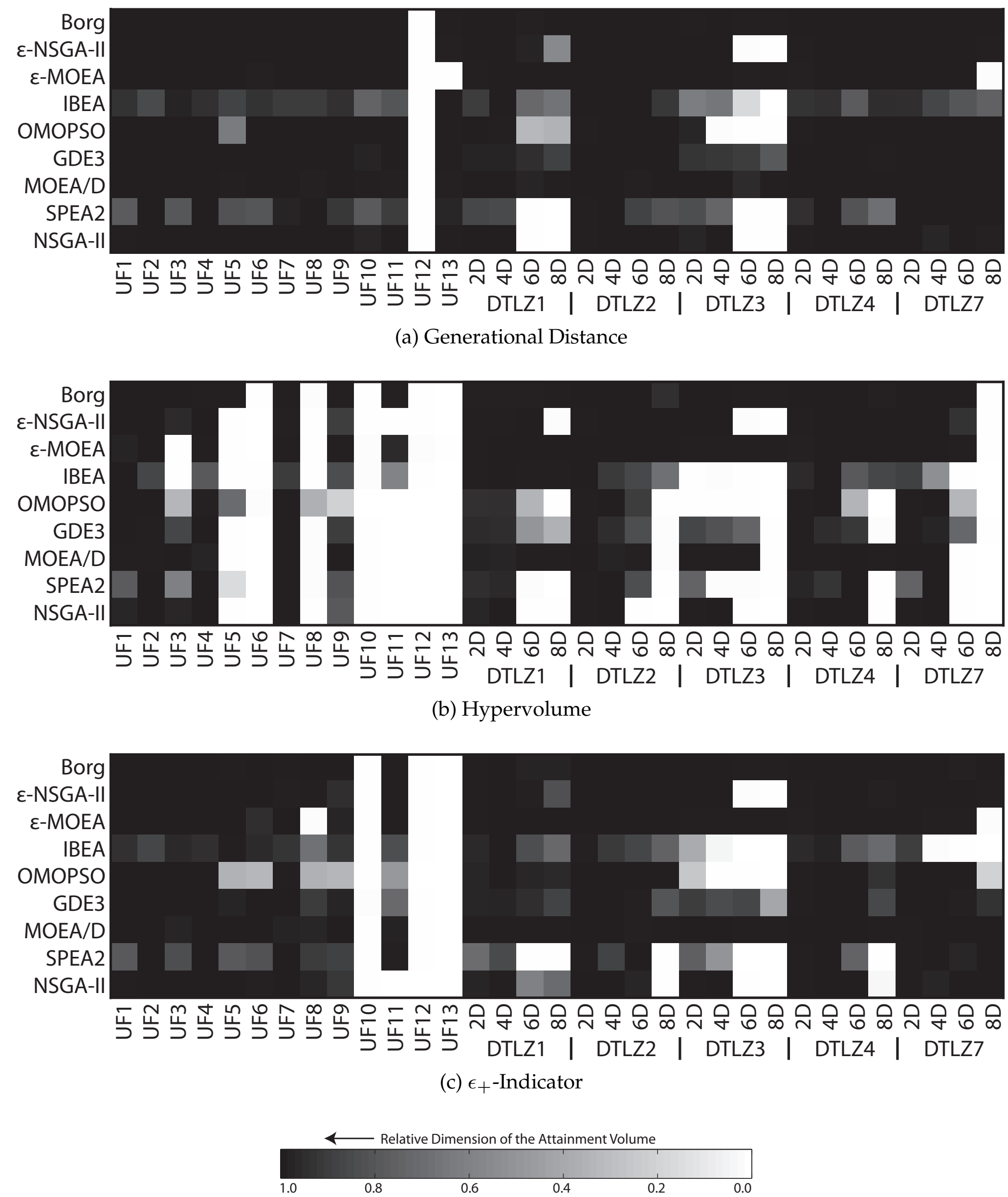


Figure 6: The efficiency of each MOEA shows the minimum number of NFE required for the algorithm to reliably (with 90% probability) produce approximation sets surpassing the 75% attainment threshold. Black regions indicate efficient algorithms requiring fewer objective function evaluations. White regions indicate cases where the algorithm failed to surpass the attainment threshold given a maximum of 1000000 evaluations.

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053



(a) Generational Distance

(b) Hypervolume

(c) ϵ_+-Indicator

Figure 7: Controllability of each algorithm on the problems studied as measured using the correlation dimension. The results are normalized such that the correlation dimensions are divided by the dimension of the hypercube used to sample each algorithm's parameter space. The correlation dimension calculation considers only those parameter sets that are able to attain the 75%-attainment threshold and consequently gives an indication of the distribution of these parameter sets in the full parametric hypervolumes sampled for each algorithm.

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

Table 5: Statistical comparison of algorithms counting the number of problems in which each MOEA was best or tied for best. The Kruskal-Wallis and Mann-Whitney U tests are used to check for statistical differences in the generational distance, hypervolume and ϵ_+-indicator values across the 50 random seed replicates. Counts are differentiated by the search control metrics: best, probability of attainment (prob), efficiency (eff) and controllability (cont).

	Hypervolume				Generational Distance				ϵ_+-Indicator			
Algorithm	Best	Prob	Eff	Cont	Best	Prob	Eff	Cont	Best	Prob	Eff	Cont
Borg	31	18	17	26	31	27	28	32	30	18	21	28
ϵ-MOEA	23	14	24	24	29	30	30	29	22	22	27	27
ϵ-NSGA-II	19	14	14	19	29	28	27	28	19	18	21	26
OMOPSO	20	15	12	10	29	24	24	25	21	16	17	16
MOEA/D	23	4	19	18	32	24	30	27	27	13	25	27
GDE3	24	8	14	7	32	21	27	23	22	11	20	15
IBEA	18	5	11	7	28	23	25	5	11	5	16	3
NSGA-II	16	8	13	13	26	21	26	25	15	9	19	18
SPEA2	16	3	9	7	26	21	24	10	13	5	13	11

important to understand their controllability. Figure 7 shows controllability, which is a measure of the spatial distribution and correlation between parameter sets in the attainment volume. Cases with low probability of attainment and high controllability signify the attainment volume forms a tightly-clustered sweet spot in a subspace of the overall parameter space. Conversely, cases with high probability of attainment and low controllability indicates the attainment volume is large but sparsely populated.

For example, compare the hypervolume values for the Borg MOEA between Figure 5 and Figure 7. Figure 5 shows the Borg MOEA has moderate attainment probabilities, but Figure 7 indicates the attainment volume is tightly clustered and forms a sweet spot. IBEA and SPEA2 show the opposite: their higher attainment probabilities correspond often with lower controllability, particularly for GD and ϵ_+-indicator. This suggests these algorithms will be more difficult to parameterize in practice, as the attainment volume is sparse. Overall, the Borg MOEA and ϵ-MOEA are the most controllable of the tested algorithms. They still struggle on several UF problems and 8D DTLZ7. ϵ-NSGA-II and MOEA/D are also strong competitors in terms of GD and ϵ_+-indicator. It is interesting to note that although Borg's multioperator search increases its parameterization requirements, its adaptive search actually serves to make the algorithm easier to use and more effective than the other algorithms on most problem instances.

Table 5 shows the number of problems each MOEA resulted in the best metric value statistically tied for the best. Ties and statistical differences were determined using a 9-way Kruskal-Wallis test preceding 2-way Mann-Whitney U tests on the results from the 50 random seed replicates using 95% confidence intervals (Sheskin, 2004). These statistical tests help guarantee that any observed differences are not a result of random chance. The MOEAs in Table 5 are shown top to bottom in the perceived ordering from best to worst. This ordering is weighted towards the hypervolume metric, as it is the strongest indicator that combines proximity and diversity into a single metric value. Across all performance measures, the Borg MOEA and ϵ-MOEA were superior on the most problems. Borg was most dominant in terms of hypervolume, whereas ϵ-MOEA was dominant on generational distance and ϵ_+-indicator. IBEA, SPEA2 and NSGA-II showed the worst performance among the tested algorithms. The large values seen in Table 5 for generational distance indicates most MOEAs were statistically indifferent from one another with respect to this metric. The wider range of values in

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

hypervolume and ϵ_+-indicator implies a number of MOEAs struggled to produce diverse approximation sets. Overall, algorithms like the Borg MOEA, ϵ-MOEA, ϵ-NSGA-II, OMOPSO and MOEA/D should be preferred in practice. Note that four of these five MOEAs include ϵ-dominance, providing experimental evidence in support of the theoretical findings of Laumanns et al. (2002).

These results combined with the statistical study performed in Hadka and Reed (2011, This Issue) helps solidify the dominance of the Borg MOEA over other state-of-the-art MOEAs. The work by Vrugt and Robinson (2007) and Vrugt et al. (2009) focusing on multimethod search supports the observation that while multimethod algorithms increase the number of algorithm parameters, the end result is a more robust and controllable tool. Nevertheless, these results show multimodal and many-objective problems still pose challenges, as is clearly observed when looking at the effectiveness and controllability of algorithms.

Now that a coarse-grained picture of search successes and failures has been established, we now explore a more fine-grained analysis of search controls using global variance decomposition. Figure 8 and Figure 9 show the first-order and interactive effects of the search parameters for the hypervolume metric for all problems. Each subplot is composed of shaded squares corresponding to the problem instance (x-axis) and the algorithm's parameters (y-axis). For the DTLZ problems, this visualization captures the change in parameter sensitivities as the objective space's dimension is increased. Black represents the most sensitive parameters whereas white identifies parameters with negligible effects. The shading corresponds to the % ensemble variance contributed by a given parameter or its interactions as identified by Sobol's global variance decomposition. Squares marked with an X indicate the bootstrap confidence intervals exceeded a window greater than +/- 20% around the expected sensitivity value (representing a 40% range), which implies the sensitivity indices could not be reliably computed. A large confidence range in the computed sensitivities is caused by the effects of parameterization not being significantly stronger than stochastic effects (i.e., low signal-to-noise). When this occurs, search is mostly independent of its parameters and is heavily influenced by purely random effects within the evolutionary algorithms. Therefore, we say the X's indicate search failure.

Note Figure 8 focuses on the Borg MOEA, ϵ-MOEA, ϵ-NSGA-II and OMOPSO, as these algorithms all share some combination of adaptive operators or ϵ-dominance archives. Figure 9 provides the sensitivities for the remaining algorithms. While these figures contains a lot of information, there are several key observations. First, for several problems there are strong first-order effects, indicating one or more parameters are independently responsible for the algorithms' performance. For Borg, ϵ-NSGA-II, ϵ-MOEA and OMOPSO, the key first-order parameter across most problems is the maximum number of evaluations. This indicates that parameterizing Borg, ϵ-NSGA-II, ϵ-MOEA and OMOPSO should prove easier in practice as the first-order impact of parameters is controlled for the most part by a single parameter, the maximum number of evaluations. Lengthening the runtime of these MOEAs will help produce better results, assuming the optimum has yet to be achieved. As a result, these algorithms should benefit from parallelization, as increasing the number of evaluations should directly result in better performance. Interestingly, these four MOEAs all utilize ϵ-dominance archives, suggesting that ϵ-dominance is an important component for controllability. Table 5 and Figure 6 also show that Borg, ϵ-NSGA-II, ϵ-MOEA and OMOPSO are in fact highly efficient on many problem instances, so it is possible to exploit their sensitivity to NFE to attain effective, reliable and efficient search.

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

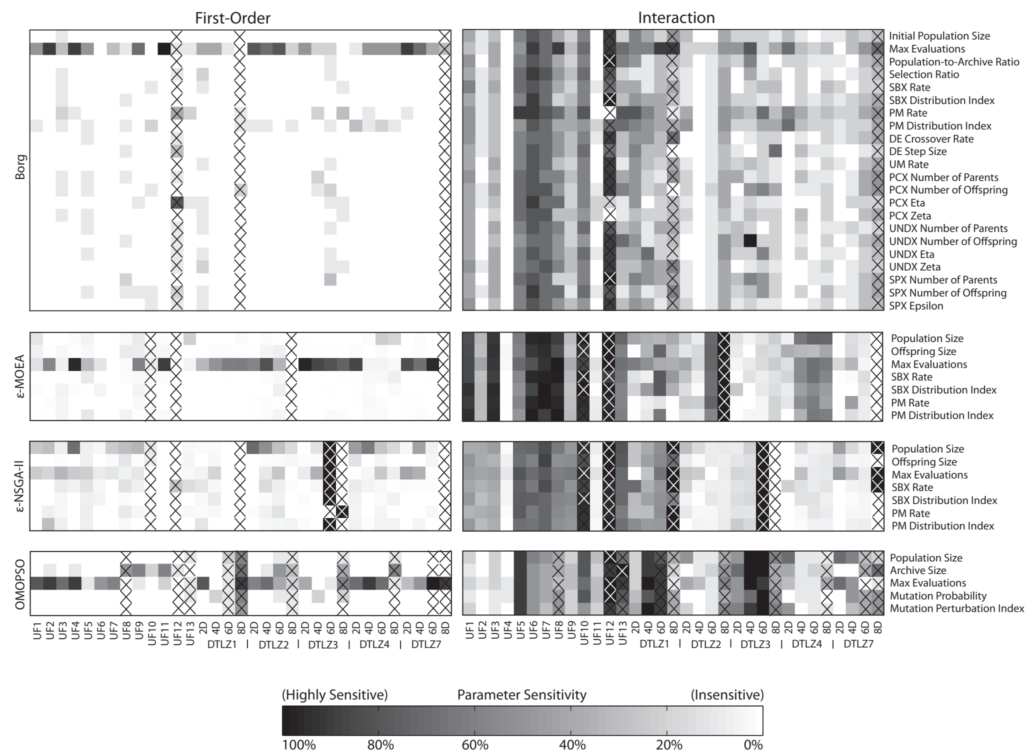


Figure 8: Sobol' sensitivities of individual algorithm parameters for all problem instances. The first-order Sobol' indices represent the single parameter contributions to the hypervolume distributions' variances. In a given problem instance, the first order indices for a given algorithm must sum to be less than or equal to 1. Interactive effects represent each parameter's contributions to the hypervolume ensembles variances through combined impacts with other parameters. Note the interactive effects do not sum to 1 for each problem dimension because each shaded cell has variance contributions that are also present in other cells (i.e., higher order interactive parametric effects). X's indicate cases when sensitivities are too uncertain to draw conclusions as determined when the bootstrap confidence intervals exceeded a window greater than +/- 20% around the expected sensitivity value.

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

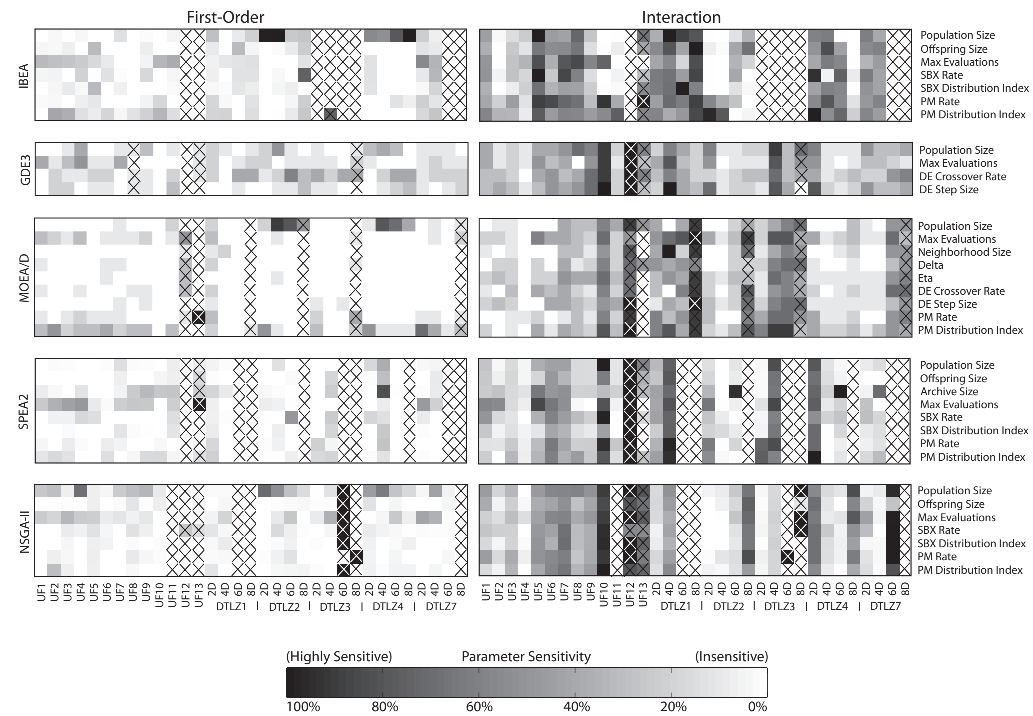


Figure 9: Sobol′ sensitivities of individual algorithm parameters for all problem instances. The first-order Sobol′ indices represent the single parameter contributions to the hypervolume distributions' variances. In a given problem instance, the first order indices for a given algorithm must sum to be less than or equal to 1. Interactive effects represent each parameter's contributions to the hypervolume ensembles variances through combined impacts with other parameters. Note the interactive effects do not sum to 1 for each problem dimension because each shaded cell has variance contributions that are also present in other cells (i.e., higher order interactive parametric effects). X's indicate cases when sensitivities are too uncertain to draw conclusions as determined when the bootstrap confidence intervals exceeded a window greater than +/- 20% around the expected sensitivity value.

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

MOEA/D and NSGA-II show strong first-order effects for population size on a number of problems. Hadka and Reed (2011, This Issue) show with control maps that these MOEAs require larger population sizes in these cases. As the algorithm runtimes grow polynomially with the population size, MOEA/D and NSGA-II are required to have long runtimes in order to maintain their performance. MOEAs not sensitive to population size will scale better in practice.

Across all tested algorithms we observe a strong trend of increasing interaction effects with increasing objective count. The level of interaction appears dependent on the problem instance, and may reflect problem difficulty. In particular, poor controllability in Figure 7 coupled with high levels of interaction between parameters indicate parameterization is difficult for a specific algorithm and problem instance. For instance, on UF11 the Borg MOEA dominates the other tested algorithms in probability of attainment and controllability, as shown in Figs. 5 and 7. This is reflected in Figure 8 in the strong first-order sensitivity to the maximum number of evaluations and weak interactive effects. On the other hand, IBEA and GDE3 show strong first-order and interactive effects spread across multiple parameters. We expect such MOEAs to be difficult to control due to the significance of many parameters. This is confirmed in Figure 7 by the weak controllability of IBEA and GDE3 in hypervolume relative to the other tested MOEAs. In this manner, a better understanding of how parameters effect search performance can be deduced from Figs. 8 and 9.

A critical concern highlighted in Figure 9 for most MOEAs that do not use ϵ-dominance archives, is how their parameter sensitivities change significantly across problem types and even within the same problem with increasing objective dimension. Moreover, their sensitivities have increasingly complex interactive dependencies for many-objective problems. Consequently, a user cannot use any "rule-of-thumb" beyond enumerative testing when using the algorithms in challenging many-objective applications, especially if they are multi-modal. These results highlight the importance auto-adaptive search frameworks such as Borg that minimize controllability challenges while maintaining efficient and reliable search.

In Hadka and Reed (2011, This Issue), we observed that for most problems, only one of Borg's recombination operators were probabilistically dominant. In other words, the auto-adaptive multi-operator approach used in Borg identified a key operator for each problem. However, Figure 8 shows that all of the operators strongly influence the overall hypervolume performance. In Vrugt and Robinson (2007) and Vrugt et al. (2009), the authors observed the same phenomenon in their multimethod approach — that while a single operator became probabilistically dominant in search, the remaining operators remained critical to the overall success of the algorithm.

6 Conclusion

Due to the increasing interest in using MOEAs to solve many-objective problems, it is necessary to understand the impact of objective scaling on search controls and failure modes. In this study, we contribute a methodology for quantifying the reliability, efficiency and controllability of MOEAs. In addition, this methodology clarifies the multivariate impacts of operator choices and parameterization on search. We have observed that many algorithms have difficulty in maintaining diverse approximation sets on problems with as few as four objectives. In addition, we have shown the necessity of diversity-maintaining archives, such as the ϵ-dominance archive, when applying MOEAs to problems with more than three objectives. A major contribution of this study is our proposed controllability measure, which permits comparing MOEAs with-

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

out arbitrary parameterization assumptions. Most algorithms are reasonably reliable, efficient and controllable for attaining approximation sets that are in close proximity to the reference sets; however, diversity is far less controllable as a problem's objective space increases in dimension. One of the major factors identified for such search failures is multimodality and the lack of ϵ-dominance archives.

Sobol's global variance decomposition was used to establish the sensitivities of each algorithm's parameters on the hypervolume of its resulting approximation set. A shift in parameter sensitivities from first-order to interactive effects was observed as the number of objectives is increased. These results can be used by researchers and practitioners when establishing parameterization guidelines. Moreover, these results suggest the need for adaptive search controls for many-objective optimization, while also indicating that adapting search controls will be non-trivial at higher problem dimensions.

The Borg MOEA's multioperator adaptivity strongly enhanced its overall effectiveness, efficiency and controllability relative to the other algorithms tested. Borg shows consistent levels of effectiveness, efficiency and controllability for a majority of the problems tested, and had very dominant performance on higher dimensional problem instances. By identifying search control issues, key parameters, and failure modes on test problems, improvements to MOEAs and their potential applicability to real-world problems can be assessed. While this study is only a first step towards understanding the impact of objective scaling on MOEAs, it has yielded several insights into the challenges faced when applying MOEAs to many-objective problems.

A Sobol's Global Variance Decomposition

Using the notation and terminology of Saltelli et al. (2008), given a square-integrable function f transforming inputs $X_1, X_2, \ldots, X_n$ into output Y,

$$Y = f(X_1, X_2, \ldots, X_n), \tag{14}$$

the global variance decomposition technique proposed by I. M. Sobol' considers the following expansion of f into terms of increasing dimension:

$$f = f_0 + \sum_i f_i + \sum_{i<j} f_{ij} + \sum_{i<j<k} f_{ijk} + \cdots + f_{ijk\ldots n}, \tag{15}$$

where each individual term is a function only over the inputs in its index (Saltelli et al., 2008; Archier et al., 1997). For example, $f_i = f_i(X_i)$ and $f_{ij} = f_{ij}(X_i, X_j)$. Sobol' proved that the individual terms can be computed using conditional expectations, such as

$$f_0 = E(Y), \tag{16}$$
$$f_i = E(Y|X_i) - f_0, \tag{17}$$
$$f_{ij} = E(Y|X_i, X_j) - f_i - f_j - f_0. \tag{18}$$

If the output Y is sensitive to input X_i, then the conditional expectation $E(Y|X_i)$ has a large variance across the values of X_i. Hence, the variance of the conditional expectation is a measure of sensitivity. The first-order effects are calculated by

$$S_i = \frac{V[f_i(X_i)]}{V[Y]} = \frac{V[E(Y|X_i)]}{V[Y]}. \tag{19}$$

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

The second-order effects are calculated with

$$S_{ij} = \frac{V[f_{ij}(X_i, X_j)]}{V[Y]} \tag{20}$$

$$= \frac{V[E(Y|X_i, X_j)]}{V[Y]} - S_i - S_j. \tag{21}$$

An important consequence of Sobol's work is the computation of total-order effects. The total effects caused by input X_i is the sum of the first-order effect S_i and all higher-order effects influenced by X_i. Thus, total-order effects are calculated by

$$S_i^T = 1 - \frac{V[E(Y|X_{\sim i})]}{V[Y]}, \tag{22}$$

where $X_{\sim i}$ represents all the inputs excluding X_i. Saltelli et al. (2008) developed the Monte Carlo technique for efficiently computing the first-, second-, and total-order effects used in this study. To validate the sensitivity results, the bootstrap technique called the *moment method* produces symmetric 95% confidence intervals, as described in Archier et al. (1997) and Tang et al. (2007). The moment method provides more reliable results with smaller resampling sizes so long as the distribution is not skewed left or right (Archier et al., 1997). We chose a resampling size of 2000 since it is both recommended in the literature and experimentally robust (Tang et al., 2007). Interested readers should refer to the cited materials for additional details.

B Acknowledgments

This work was supported in part through instrumentation funded by the National Science Foundation through grant OCI-0821527.

References

Adra, S. F. and Fleming, P. J. (2009). A diversity management operator for evolutionary many-objective optimisation. In *Evolutionary Multi-Criterion Optimization*, pages 81–94, Berlin / Heidelberg. Springer.

Archier, G., Saltelli, A., and Sobol, I. (1997). Sensitivity measures, anova-like techniques and the use of bootstrap. *Journal of Statistical Computation and Simulation*, 58:99–120.

Baker, G. L. and Gollub, J. P. (1990). *Chaotic dynamics : an introduction*. Cambridge University Press, Cambridge.

Bosman, P. A. and Thierens, D. (2003). The balance between proximity and diversity in multiobjective evolutionary algorithms. *IEEE Transactions on Evolutionary Computation*, 7(2):174–188.

Brockhoff, D., Friedrich, T., Hebbinghaus, N., Klein, C., Neumann, F., and Zitzler, E. (2007). Do additional objectives make a problem harder? In *Genetic and Evolutionary Computation Conference (GECCO 2007)*, pages 765–772, New York, NY, USA. ACM Press.

Coello Coello, C. A., Lamont, G. B., and Van Veldhuizen, D. A. (2007). *Evolutionary Algorithms for Solving Multi-Objective Problems*. Springer Science+Business Media, LLC, New York, NY, USA.

Corne, D. W. and Knowles, J. D. (2000). The Pareto-envelope based selection algorithm for multiobjective optimization. In *Parallel Problem Solving from Nature (PPSN VI)*, pages 839–848, London, UK. Springer-Verlag.

Corne, D. W. and Knowles, J. D. (2007). Techniques for highly multiobjective optimisation: some nondominated points are better than others. In *Genetic and Evolutionary Computation Conference (GECCO 2007)*, pages 773–780, New York, NY, USA. ACM.

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

Deb, K. and Jain, S. (2002). Running performance metrics for evolutionary multi-objective optimization. KanGAL Report No. 2002004, Kanpur Genetic Algorithms Laboratory (KanGAL).

Deb, K., Mohan, M., and Mishra, S. (2003). A fast multi-objective evolutionary algorithm for finding well-spread Pareto-optimal solutions. KanGAL Report No. 2003002, Kanpur Genetic Algorithms Laboratory (KanGAL).

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2000). A fast elitist multi-objective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197.

Deb, K. and Saxena, D. K. (2006). Searching for Pareto-optimal solutions through dimensionality reduction for certain large-dimensional multi-objective optimization problems. In *The 2006 IEEE Congress on Evolutionary Computation*, pages 3353–3360.

Deb, K., Thiele, L., Laumanns, M., and Zitzler, E. (2001). Scalable test problems for evolutionary multi-objective optimization. TIK-Technical Report No. 112, Computer Engineering and Networks Laboratory (TIK), Swiss Federal Institute of Technology (ETH).

Deb, K., Thiele, L., Laumanns, M., and Zitzler, E. (2002). Scalable multi-objective optimization test problems. In *Congress on Evolutionary Computation (CEC 2002)*, pages 825–830.

di Pierro, F. (2006). *Many-Objective Evolutionary Algorithms and Applications to Water Resources Engineering*. PhD thesis, University of Exeter, UK.

di Pierro, F., Khu, S.-T., and Savic, D. A. (2007). An investigation on preference order ranking scheme for multiobjective evolutionary optimization. *IEEE Transactions on Evolutionary Computation*, 11(1):17–45.

Drechsler, N., Drechsler, R., and Becker, B. (2001). Multi-objective optimisation based on relation favour. In *Evolutionary Multi-Criterion Optimization*, pages 154–166, Berlin / Heidelberg. Springer.

Edwards, A. L. (1993). *An Introduction to Linear Regression and Correlation*. W. H. Freeman and Co, San Francisco.

Farina, M. and Amato, P. (2004). A fuzzy definition of "optimality" for many-criteria optimization problems. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 34(3):315–326.

Ferringer, M. P., Spencer, D. B., and Reed, P. (2009). Many-objective reconfiguration of operational satellite constellations with the large-cluster epsilon Non-dominated Sorting Genetic Algorithm-II. In *IEEE Congress on Evolutionary Computation (CEC 2009)*, pages 340–349.

Fleming, P. J., Purshouse, R. C., and Lygoe, R. J. (2005). Many-objective optimization: An engineering design perspective. In *Evolutionary Multi-Criterion Optimization*, Lecture Notes in Computer Science, pages 14–32. Springer, Berlin / Heidelberg.

Fonseca, C. M. and Fleming, P. J. (1993). Genetic algorithms for multiobjective optimization: Formulation, discussion and generalization. In *Genetic Algorithms: Proceedings of the Fifth International Conference*, pages 416–423, San Mateo, CA, USA. Morgan Kaufmann.

Fonseca, C. M. and Fleming, P. J. (1998). Multiobjective optimization and multiple constraint handling with evolutionary algorithms — part 1: A unified formulation. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 28(1):26–37.

Goh, C.-K. and Tan, K. C. (2009). *Evolutionary Multi-objective Optimization in uncertain Environments*. Springer, Berlin.

Goldberg, D. E. (1989a). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Publishing Company, Reading, MA, USA.

Goldberg, D. E. (1989b). Sizing populations for serial and parallel genetic algorithms. In *3rd International Conference on Genetic Algorithms*, pages 70–79.

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

Grassberger, P. and Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica D Nonlinear Phenomena*, 9:189–208.

Hadka, D. and Reed, P. (2011). Borg: An auto-adaptive many-objective evolutionary computing framework. *Evolutionary Computation*.

Harik, G. R. and Lobo, F. G. (1999). A parameter-less genetic algorithm. *IEEE Transactions on Evolutionary Computation*, pages 523–528.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI, USA.

Horn, J. and Nafpliotis, N. (1993). Multiobjective optimization using the Niched Pareto Genetic Algorithm. Technical Report 93005, University of Illinois.

Huband, S., Hingston, P., Barone, L., and While, L. (2006). A review of multiobjective test problems and a scalable test problem toolkit. *IEEE Transactions on Evolutionary Computation*, 10(5):477–506.

Hughes, E. J. (2005). Evolutionary many-objective optimisation: many once or one many? In *The 2005 IEEE Congress on Evolutionary Computation*, volume 1, pages 222–227.

Iorio, A. and Li, X. (2008). Improving the performance and scalability of differential evolution. In *Simulated Evolution and Learning*, volume 5361 of *Lecture Notes in Computer Science*, pages 131–140. Springer Berlin / Heidelberg.

Ishibuchi, H., Tsukamoto, N., Hitotsuyanagi, Y., and Nojima, Y. (2008a). Effectiveness of scalability improvement attempts on the performance of NSGA-II for many-objective problems. In *Genetic and Evolutionary Computation Conference (GECCO 2008)*, pages 649–656, New York, NY, USA. ACM.

Ishibuchi, H., Tsukamoto, N., and Nojima, Y. (2008b). Behavior of evolutionary many-objective optimization. In *Tenth International Conference on Computer Modeling and Simulation (UKSIM 2008)*, pages 266–271.

Ishibuchi, H., Tsukamoto, N., Sakane, Y., and Nojima, Y. (2010). Indicator-based evolutionary algorithm with hypervolume approximation by achievement scalarizing functions. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, GECCO '10, pages 527–534.

Kasprzyk, J., Reed, P., Kirsch, B., and Characklis, G. (2011). Many-objective de novo water supply portfolio planning under deep uncertainty. *Environmental Modelling & Software*. doi:10.1016/j.envsoft.2011.04.003.

Kasprzyk, J. R., Reed, P. M., Kirsch, B. R., and Characklis, G. W. (2009). Managing population and drought risks using many-objective water portfolio planning under uncertainty. *Water Resources Research*, 45. W12401, doi:10.1029/2009WR008121.

Khare, V., Yao, X., and Deb, K. (2003). Performance scaling of multi-objective evolutionary algorithms. In *Evolutionary Multi-Criterion Optimization*, pages 376–390, Berlin / Heidelberg. Springer.

Knowles, J. and Corne, D. (2002). On metrics for comparing non-dominated sets. In *Congress on Evolutionary Computation (CEC 2002)*, volume 1, pages 711–716.

Knowles, J. and Corne, D. (2007). Quantifying the effects of objective space dimension in evolutionary multiobjective optimization. In *Evolutionary Multi-Criterion Optimization*, Lecture Notes in Computer Science, pages 757–771, Berlin / Heidelberg. Springer.

Knowles, J. D. and Corne, D. W. (1999). Approximating the nondominated front using the Pareto Archived Evolution Strategy. *Evolutionary Computation*, 8:149–172.

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

Kollat, J., Reed, P., and Maxwell, R. (2011). Many-objective groundwater monitoring network design using bias-aware ensemble kalman filtering, evolutionary optimization, and visual analytics. *Water Resources Research*, 47. W02529, doi:10.1029/2010WR009194.

Kollat, J. B. and Reed, P. M. (2006). Comparison of multi-objective evolutionary algorithms for long-term monitoring design. *Advances in Water Resources*, 29(6):792–807.

Kollat, J. B. and Reed, P. M. (2007). A computational scaling analysis of multiobjective evolutionary algorithms in long-term groundwater monitoring applications. *Advances in Water Resources*, 30(3):408–419.

Kukkonen, S. and Deb, K. (2006). A fast and effective method for pruning of non-dominated solutions in many-objective problems. In *Parallel Problem Solving from Nature (PPSN IX)*, Lecture Notes in Computer Science, pages 553–562, Berlin / Heidelberg. Springer.

Kukkonen, S. and Lampinen, J. (2005). GDE3: the third evolution step of generalized differential evolution. In *The 2005 IEEE Congress on Evolutionary Computation*, volume 1, pages 443–450.

Laumanns, M., Thiele, L., Deb, K., and Zitzler, E. (2002). Combining convergence and diversity in evolutionary multi-objective optimization. *Evolutionary Computation*, 10(3).

Nayfeh, A. H. and Balachandran, B. (1995). *Applied Nonlinear Dynamics: Analytical, Computational, and Experimental Methods*. Wiley Inter-Science.

Praditwong, K. and Yao, X. (2007). How well do multi-objective evolutionary algorithms scale to large problems. In *Congress on Evolutionary Computation (CEC 2007)*, pages 3959–3966.

Purshouse, R. C. and Fleming, P. J. (2003). Evolutionary many-objective optimisation: an exploratory analysis. In *Congress on Evolutionary Computation (CEC 2003)*, volume 3, pages 2066–2073.

Purshouse, R. C. and Fleming, P. J. (2007). On the evolutionary optimization of many conflicting objectives. *IEEE Transactions on Evolutionary Computation*, 11(6):770–784.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global Sensitivity Analysis: The Primer*. Wiley.

Schaffer, D. J. (1984). *Multiple Objective Optimization with Vector Evaluated Genetic Algorithms*. PhD thesis, Vanderbilt University.

Sheskin, D. J. (2004). *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC.

Sierra, M. R. and Coello Coello, C. A. (2005). Improving PSO-based multi-objective optimization using crowding, mutation and ϵ-dominance. In *Evolutionary Multi-Criterion Optimization*, Lecture Notes in Computer Science, pages 505–519, Berlin / Heidelberg. Springer.

Srinivas, N. and Deb, K. (1993). Multiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary Computation*, 2(3):221–248.

Srivastava, R. P. (2002). Time continuation in genetic algorithms. Technical report, Illinois Genetic Algorithm Laboratory.

Sülflow, A., Drechsler, N., and Drechsler, R. (2007). Robust multi-objective optimization in high dimensional spaces. In *Evolutionary Multi-Criterion Optimization*, pages 715–726, Berlin / Heidelberg. Springer.

Tang, Y., Reed, P., Wagener, T., and van Werkhoven, K. (2007). Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation. *Hydrology and Earth System Sciences*, 11(2):793–817.

Teytaud, O. (2006). How entropy-theorems can show that on-line approximating high-dim Pareto-fronts is too hard. In *PPSN BTP Workshop*.

Evolutionary Computation Just Accepted MS.
doi:10.1162/EVCO_a_00053

Teytaud, O. (2007). On the hardness of offline multi-objective optimization. *Evolutionary Computation*, 15(4):475–491.

Vrugt, J. A. and Robinson, B. A. (2007). Improved evolutionary optimization from genetically adaptive multimethod search. *Proceedings of the National Academy of Sciences*, 104(3):708–711.

Vrugt, J. A., Robinson, B. A., and Hyman, J. M. (2009). Self-adaptive multimethod search for global optimization in real-parameter spaces. *IEEE Transactions on Evolutionary Computation*, 13(2):243–259.

Wagner, T., Beume, N., and Naujoks, B. (2007). Pareto-, aggregation-, and indicator-based methods in many-objective optimization. In *Evolutionary Multi-Criterion Optimization*, pages 742–756, Berlin / Heidelberg. Springer.

Zhang, Q., Liu, W., and Li, H. (2009a). The performance of a new version of MOEA/D on CEC09 unconstrained MOP test instances. In *Congress on Evolutionary Computation (CEC 2009)*, pages 203–208, Piscataway, NJ, USA. IEEE Press.

Zhang, Q. and Suganthan, P. N. (2009). Final report on CEC'09 moea competition. In *Congress on Evolutionary Computation (CEC 2009)*, Piscataway, NJ, USA. IEEE Press.

Zhang, Q., Zhou, A., Zhao, S., Suganthan, P. N., Liu, W., and Tiwari, S. (2009b). Multiobjective optimization test instances for the CEC 2009 special session and competition. Technical Report CES-487, University of Essex.

Zitzler, E. and Künzli, S. (2004). Indicator-based selection in multiobjective search. In *Parallel Problem Solving from Nature (PPSN VIII)*, Lecture Notes in Computer Science, pages 832–842, Berlin / Heidelberg. Springer.

Zitzler, E., Laumanns, M., and Thiele, L. (2002a). *SPEA2: Improving the Strength Pareto Evolutionary Algorithm For Multiobjective Optimization*. CIMNE, Barcelona, Spain.

Zitzler, E., Laumanns, M., Thiele, L., Fonseca, C. M., and da Fonseca, V. G. (2002b). Why quality assessment of multiobjective optimizers is difficult. In *Genetic and Evolutionary Computation Conference (GECCO 2002)*, pages 666–674. Morgan Kaufmann Publishers.

Zitzler, E. and Thiele, L. (1999). Multiobjective evolutionary algorithms: A comparative case study and the strength Pareto approach. *IEEE Transactions on Evolutionary Computation*, 3(4):257–271.

Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C. M., and da Fonseca, V. G. (2002c). Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on Evolutionary Computation*, 7(2):117–132.