Journal of the Operational Research Society (2016) 67, 676–689 © 2016 Operational Research Society Ltd. All rights reserved. 0160-5682/16

www.palgrave-journals.com/jors/

Good Laboratory Practice for optimization research

Graham Kendall[1,2*], Ruibin Bai[3], Jacek Błazewicz[4], Patrick De Causmaecker[5], Michel Gendreau[6], Robert John[1], Jiawei Li[1], Barry McCollum[7], Erwin Pesch[8], Rong Qu[1], Nasser Sabar[2], Greet Vanden Berghe[9] and Angelina Yee[2]

[1]University of Nottingham, Nottingham, UK; [2]University of Nottingham Malaysia Campus, Semenyih, Malaysia; [3]University of Nottingham Ningbo, Ningbo, China; [4]Poznan University of Technology, Poznan, Poland; [5]KU Leuven, Campus Kulak, Kortrijk, Belgium; [6]University of Montreal, Montreal, Canada; [7]Queen's University Belfast, Belfast, UK; [8]Universität Siegen, Siegen, Germany; and [9]KU Leuven, Technology Campus Gent, Gent, Belgium

Good Laboratory Practice has been a part of non-clinical research for over 40 years. Optimization Research, despite having many papers discussing standards being published over the same period of time, has yet to embrace standards that underpin its research. In this paper we argue the need to adopt standards in optimization research. Building on previous papers, many of which have suggested that the optimization research community should adopt certain standards, we suggest a concrete set of recommendations that the community should adopt. We also discuss how the proposals in this paper could be progressed.

Journal of the Operational Research Society (2016) **67**(4), 676–689. doi:10.1057/jors.2015.77
Published online 21 October 2015

Keywords: optimization; operations research; reproducibility

1. Introduction

As far back as Ignizio (1971), and perhaps before, there was discussion in the scientific literature as to how the scientific community should report algorithmic results. Looking at some recent papers, little seems to have changed, and if we compare the optimization research community against other communities (eg, non-clinical research) it is debatable whether it would stand up to close scrutiny with regard to the research methods that are employed. That is not to say that the research is not conducted without relevant integrity, but the scientific process, and the way that experiments are reported, could be subject to some criticism. It should also be noted that other disciplines are not without criticism (Anon, 2013).

The academic community is also aware that if we do archive our data and results, then we run the risk of losing this data. The study in Vines *et al* (2013) shows that the availability of data in research papers is strongly affected by the age of the article. They conclude that data cannot be reliably preserved by individual researchers. The optimization research community should ensure that the data that underpins its research findings is available in a centralized location for future researchers, and we should not be reliant on single researchers or research groups.

Good Laboratory Practice (GLP) has been established for over 40 years in non-clinical research and is used to assess the safety or efficacy of chemicals, including pharmaceuticals, to man, animals and the environment. It provides a quality system of management controls for research laboratories and organizations to ensure the uniformity, consistency, reliability, reproducibility, quality and integrity of chemical research. Bioinformatics is an area where some of these practices have spread to optimization research, due to the close relationship between biologists and optimization research (Błazewicz *et al*, 2005; Lukasiak *et al*, 2010). As an example, Bioinformatics insists on computer code being made available, which is one of the areas discussed in this paper (see Section 4.12).

Given its long history (Gass and Assad, 2006), dating back to positioning radar stations in World War II (Brown, 1999), it is, perhaps, surprising that optimization research has few documented quality standards that underpin at least some of its research. Non-clinical research, undoubtedly, has the potential to affect the lives of a great number of people, whether positively or adversely. The same is also true of optimization research. Perhaps even more so considering the scope of its reach. There are some *classic* areas which optimization researchers have embraced, and regularly report on. To give just a small sample, we provide this short list of a few representative optimization areas. We have identified papers from the following areas, both recent and from over 50 years ago, just to demonstrate how established some of them are.

- *Assignment problems*—Cattrysse and Van Wassenhove (1992), Loiola *et al* (2007), Pentico (2007)
- *Bioinformatics*—Błazewicz *et al* (2005), Lukasiak *et al* (2010)
- *Forecasting*—Fildes (1979, 1985), Fildes *et al* (2008), Ghobbar and Friend (2003), Trapero *et al* (2015)

*Correspondence: Graham Kendall, University of Nottingham Malaysia Campus, Jalan Broga, 43500 Semenyih, Selangor, Malaysia.

- *Job shop scheduling*—Kiran and Smith (1984), Błazewicz *et al* (1996, 2007), Cheng *et al* (1999a, b), Brucker *et al* (2007)
- *Knapsack problem*—Salkin and Kluyver (1975), Lin (1998), Pisinger (2007), Wilbaut *et al* (2008), Lust and Teghem (2012)
- *Personnel scheduling*—Miller (1976), Ernst *et al* (2004)
- *Sports scheduling*—Rasmussen and Trick (2008), Kendall *et al* (2010)
- *Traveling salesman problem (TSP)*—Bellmore and Nemhauser (1968), Burkard *et al* (1998), Lust and Teghem (2010), Cook (2011)
- *Vehicle routing*—Cordeau *et al* (2002), Braysy *et al* (2004), Braysy and Gendreau (2005a, b), Archetti and Speranza (2008), Gendreau *et al* (2008), Jaillet and Wagner (2008), Laporte (2009), Doerner and Schmid (2010), Vidal *et al* (2013)

Of course, there are many other areas that also come under the optimization research radar. Indeed, it is difficult to think of a discipline/sector that either does not currently use optimization research, or could benefit from some form of optimization. Any sector that ever use phrases such as *minimize, maximize, reduce cost, increase profit, make more efficient, save time/money/energy* and so on can benefit from the tools and methodologies available to optimization researchers.

Given the potential high impact and visibility of its research it is beholden on the community to ensure that the research it carries out is done to the highest possible standards, which would stand up to external scrutiny should that be required. There is no suggestion that any research that is currently conducted falls short of these standards. Indeed, as far as the authors are aware, our research is well respected and the peer review process is fair, stringent and works well.

In this paper, we propose a set of procedures that we hope will further enhance the quality of the research that the optimization research community carries out. If these procedures were adopted, it would ensure that different laboratories/researchers were working to the same minimum standards which would enhance the reproducibility of results, the comparison of results, and the efficiency of individual researchers and teams. Moreover, we suggest that papers that adopt the standards presented in this paper would be able to identify this in the paper, which would provide a further indication of the paper's quality.

The recommendations in this paper are applicable to most, if not all, areas of optimization, where simulations are carried out, whether that is exact methods, heuristics, meta-heuristics or hyper-heuristics. These algorithms, especially where multiple computational experiments are required, often require some form of statistical analysis, so we need to ensure that this element of the algorithm design/reporting is handled appropriately. Many domains also have benchmark data sets, providing opportunities for further standards to be adopted.

For completeness, we provide the following definitions of the main areas that are covered by the recommendations in this paper. These definitions are mostly drawn from the recent scientific literature and are representative, rather than being definitive, as there are no widely accepted definitions for these terms.

Exact optimization methods (Jourdan *et al*, 2009): 'Exact methods find the optimal solution and assess its optimality'.

Heuristics (Zanakis *et al*, 1989): 'Their [heuristics] appeal stems from their ability to produce quickly near-optimal solutions to difficult optimization problems. As opposed to the advanced mathematics that is required to develop theoretical results in the optimization arena, the development of heuristics is chiefly an art and a creative problem solving endeavor' and 'Construction algorithms generate a solution by adding individual components (eg, nodes, arcs, variables) one at a time until a feasible solution is obtained. Greedy' algorithms, seeking to maximize improvement at each step, comprise a large class of construction heuristics. In most construction heuristics, a feasible solution is not found until the end of the procedure' and 'Improvement heuristics begin with a feasible solution and successively improve it by a sequence of exchanges or mergers in a local search. Generally, a feasible solution is maintained throughout the procedure'.

Metaheuristics (Talbi, 2009; Sörensen and Glover, 2013): 'A metaheuristic is a high-level problem-independent algorithmic framework that provides a set of guidelines or strategies to develop heuristic optimization algorithms. The term is also used to refer to a problem-specific implementation of a heuristic optimization algorithm according to the guidelines expressed in such a framework'.

Hyper-heuristics (Burke *et al*, 2013): 'A search method or learning mechanism for selecting or generating heuristics to solve computational search problems'.

In this paper, we provide a set of suggested standards that optimization laboratories/researchers might want to adopt. The recommendations may require some refining, which is why we have provided a suggested name (Good Laboratory Practice for Optimization Research (GLP4OPT)), along with a version control mechanism so that it is clear what standards are being adopted at any given time. The version we are proposing in this paper is version *GLP4OPT version 1.00*.

2. Related work

2.1. Optimization research

One of the earliest papers that looked at comparing computational methods can be found in Hoffman *et al* (1953), where they compared several types of linear programme solvers, concluding that the simplex method was the best from three examined. It is interesting to note that some of the discussion was about preprocessing from magnetic tapes and the effect this has on the overall computation time. We do not have to consider these type of devices today, but it is interesting to note that researchers hardly ever mention the overheads of certain devices these days.

In Ignizio (1971) there were calls for establishing standards for comparing algorithms. Ignizio proposed a standard reporting format and measurement standards. In its summary, the author says 'This author believes that our profession needs to establish such standards as previously discussed. By establishing such standards we can expect to (1) upgrade the status of our profession, (2) eliminate many marginal publications of dubious value, and (3) provide a means by which algorithms may be selected more objectively by the practitioner'. We would suggest that over 40 years on, that this is still an aspiration for the optimization research community.

Another call was made for guidelines in 1978 by Jackson and Mulvey. They said that 'Unfortunately, the methodology for conducting such computational experiments has not received systematic study, and no set of generally accepted guidelines has been available'. They surveyed 50 papers, reporting how they have reported their experimental results. Their concluding remarks include: 'This paper shows that a consensus for conducting computational experiments has not been reached, although patterns can be detected within certain areas of mathematical programs. On the whole, the more recent experiments appear improved in methodology only slightly over their predecessors'. Over 30 years later Sörensen (2015) makes similar comments with regard to taking up guidelines—'By setting up a statistical experiment, the main and interaction effects of the different algorithmic parameters on the solution quality produced by the meta-heuristic can be determined in a statistically valid way, and the optimal combination of parameter levels can be determined. Methods and guidelines to perform this step in the algorithm design are readily available (Coy et al, 2001; Adenso-Diaz and Laguna, 2006), but have not caught on in any significant way'.

In August 1973, this topic was discussed by the Mathematical Programming community when it was raised at the Stanford Mathematical Programming Symposium (Balinski, 1978). In an editorial note, Balinski (1978) noted that the paper by Crowder *et al* (1978) would promote further discussion, which might eventually lead to a formal position being taken by the Committee on Algorithms. Balinski further notes that flexible editorial policy should still be used and any guidelines should not be used as a way to simply reject a paper.

The paper referred to by Balinski (Crowder *et al*, 1978) presents a set of recommended standards for presenting computational experiments from mathematical programming. Their opening remarks state that a review of papers in the area show a high level of mathematical expertise from the authors but this does not translate into the rigor that would be expected in order to reproduce the experiments being reported, for example, important parameters are often omitted. Crowder *et al* further say that the mathematical programming community does not compare well to social science research in their scientific rigor. They state that part of the problem is that there are no published standards and see their paper going some way toward resolving this issue. The paper specifically discusses experimental design, reporting of computational results and suggestions such as how to present an algorithm. They conclude the paper with a checklist of important points.

While recognizing that it is almost impossible to totally reproduce a given experiment, due to technological changes, it should be possible to produce the results within some tolerance, recognizing the technological changes that have taken place. Crowder *et al* further note that 'It is here where editors are required to exercise their judgment on whether sufficient evidence has been provided, to convince them that the criterion of reproduceability could be met if tested'. Moreover, the authors stress, 'However, an absolute, reasonable, and scientifically justifiable criterion should be that the authors themselves be able to replicate their experiment. This is one of the basic principles of the scientific method and should be actively pursued in any properly conducted scientific inquiry'. It is also interesting to note that they suggest that two versions of the paper should be submitted. The first will be published in the journal. The second provides all the details that may not be appropriate for the published article. Given that the internet has advanced so much since 1978, it would be much easier to meet this aspiration by the use of supplementary files.

The same authors published a related paper the following year (Crowder *et al*, 1979), which covered a lot of similar material, including a checklist of important points to consider when evaluating or reporting computational experiments. These were split into recommended and optional points, and included presenting a complete description of the algorithm; the programming language and compiler used, along with the options that were set; the computer environment; a description of any special data structures that were utilized; a description of the input data; the use of standard test problems and reporting of the results; a clear statement of the objectives of the experiment which can be evaluated by the reviewers and a complete description of the problem generator.

Barr *et al* (1995), in the first issue of the *Journal of Heuristics*, discussed how to design and report computational experiments. This is a thought provoking paper, presenting some historical readings which discuss how mathematical programming methodologies should be compared (Crowder *et al*, 1979). In our view, this paper is essential reading for anybody who has an interest in this area. Barr *et al*'s paper could be seen as being critical of the heuristics/meta-heuristics community. It was actually providing some suggestions as to how the community could be more rigorous in reporting its experiments and findings.

Jackson *et al* (1991) say that 'Controversy often surrounds the reporting of results from scientific experimentation. Subtle points with unrecognized but profound effects are sometimes overlooked, and testing procedures are sometimes inadequate'. Their report does not propose a new set of guidelines but rather 'We have two goals: (1) to present a concise set of principles to guide authors, editors, and referees in assessing computational tests, and (2) to clarify existing guidelines in certain areas'. Their report builds on the work of Crowder *et al* (1979), using this as the base set of guidelines. The authors say that it is not

possible to provide a complete set of guidelines but, instead, to provide a set of principles. These being, '(1) the results presented must be sufficient to justify the claims made; (2) there must be sufficient detail to allow reproducibility of the results; and (3) it is not the referee's duty to reproduce the results'.

Golden and Stewart (1985) consider ways that statistical analysis can be carried out on the TSPs. Although focusing on the TSP (as that is the subject of the book from which this chapter is taken from) the discussion on statistical tests and how to apply them is valuable. A worked example of applying the Wilcoxon signed rank test and the Friedman test demonstrates how three heuristics can be compared, rather than just making an assumption that one of the heuristic dominates the other two. Limitations of using these statistical tests are given, suggesting that an *expected utility function* is used instead. This is simple to apply but does rely on some arbitrary assumptions.

The following year Golden *et al* (1986) published an article that looked at the experimentation issues in optimization. The paper focuses on three case studies, with an initial discussion on the statistical analysis. It is interesting to note that one of the case studies considers simulated annealing (Kirkpatrick *et al*, 1983), which is one of the first times that a meta-heuristic is discussed in this context, due to its (then) recent introduction.

Greenberg (1990) considers how and why to perform computational testing and how much testing should be conducted. The paper also addresses the issue of when commercial confidentiality is claimed. The author's view is that if a research contribution is being claimed then 'Any research paper whose merit depends critically (perhaps decisively) on the computational results must be prepared to have the results reviewed by referees'. Later, this point is reemphasized—'Under no circumstances will a paper be published in the *ORSA Journal on Computing* whose merit and information content depend critically on empirical claims that are not subjected to such review'.

Greenberg (1990) was written from the perspective of a specific journal. *IIE Transactions* also published its own guidelines when reporting computational results (Lee *et al*, 1993). Topics covered include comparison with existing methods, performance analysis, reproducibility and complexity analysis.

Hooker (1994, 1995) wrote two articles in the mid-nineties. In his 1994 paper, he called for the need to have an empirical science of algorithms. He argues for rigorous experimental design and analysis and also for the development of empirically-based explanation theories. An analytical approach to algorithm analysis has, he argues, turned into a science, whereas empirical testing has not. His concerns are expressed as follows: 'Computational experiments are widely reported in scholarly publications. But these efforts fall short of science on several levels. To begin, the testing is usually quite informal, at least in the optimization research literature. One occasionally sees tests conducted according to the principles of experimental design, or results analyzed using rigorous statistical methods. But only occasionally. Not even minimal standards of reproducibility are observed by most authors'. He goes onto to say that as a discipline, we do not encourage the publishing of negative results, which could be as important as a positive result when testing a given hypothesis. In his 1995 paper, Hooker argues that a more scientific approach is required in controlled experimentation. He likens the current state of play as resembling a track meet, where the aim is simply to beat the opponent. If the selected algorithm beats all other algorithms on selected test cases, it is submitted for publication, else the research is written off as a failure and the next idea is attempted. One section of the paper is devoted to the 'evils of competitive testing', outlining the problems it presents (eg, different levels of coding skills, different machines being used, the problems of using benchmark instances, etc). He goes onto say how all these issues can be removed, but it would need support and recognition from the scientific community.

McGeoch (1996) also makes the case for a more scientific way to study algorithms. One-dimensional bin packing is used as an example application, in order to discuss the issues addressed in the paper. Issues specifically addressed include planning the experiments, conducting a pilot study, developing and running experiments and statistics and data analysis.

Ahuja and Orlin (1996) discuss the use of CPU time as a measure of performance. They suggest that this is too implementation dependent and suggest that researchers should use *representative operation counts*, being a fairer way to provide comparative analysis.

Rardin and Uzsoy (2001) provide a tutorial of carrying out heuristic optimization experiments. This paper is also essential reading for those interested in this area. The paper focuses 'squarely on empirical evaluation of heuristic optimization methods'. They acknowledge that 'the questions are difficult, and there are no clear right answers', seeking 'only to highlight the main issues and present a number of alternative ways of addressing them under different circumstances, as well as a number of pitfalls to avoid'. Among the topics discussed in the paper are research *versus* development, designing computational experiments, sources of test instances (real world, random variants of real world instances, libraries of instances, random generated instances) and measuring performance, carrying out analysis and presenting results. They also present a case study, based on one machine scheduling.

A more recent paper Sörensen (2015) is certainly critical of this community. In his opening, the author says: 'In this paper, we will argue that this line of research [the proliferation of meta-heuristic methodologies] is threatening to lead the area of meta-heuristics away from scientific rigor'. In the concluding remarks, it states 'Although several authors have developed procedures to make a statistically sound comparison (see, eg, Barr et al, 1995; Rardin and Uzsoy, 2001), widespread acceptance of such procedures is lacking. Perhaps a set of tools is needed, ie, a collection of statistical programs or libraries specifically designed to determine the relative quality of a set of algorithms on a set of problem instances. These should both be easy to use, and their results should be easy to interpret.

Until such tools are available and a specific comparison protocol is enforced by journal editors and reviewers, the door is left open for researchers to select the method of comparison that proves the point it is intended to prove'. A related theme can be found in Hooker (2007) who argues that a bad future for Constraint Programming and *Operations Research* is one 'which is defined by the computational techniques, rather than by the phenomena they study', later saying that operations research has remained strong for the past 50 years despite the fact that many new methodologies have been introduced.

A very recent paper (Boylan *et al*, 2015) investigated replicability and reproducibility in forecasting. The study took an important paper (Miller and Williams, 2003), which had won an outstanding paper award and is highly cited, and asked two independent research teams to replicate the results. They encountered several difficulties, including data clarification, method clarification, using different software and accuracy measures. The two teams reached almost the same results as the original paper, but they were different and the teams would not have arrived at the same conclusions as Miller and Williams.

We actually started writing this paper before being aware of papers such as Rardin and Uzsoy (2001), Sörensen (2015) and Boylan *et al* (2015), but if we had read those papers before conducting our literature survey, it would certainly have motivated us to write this paper. Many of the ideas expressed by the authors we cite above are captured in the ideas we present later in this paper, by way of a number of recommendations.

We have gone one stage further than previous authors by providing an explicit set of recommendations that we hope will form the basis as a set of guidelines for the community to adopt. Non-clinical researchers abide by the GLP guidelines. The meta-heuristic community, we believe, needs such practices, many of which have been presented by other authors, but there is a need for the community to embrace and use them, rather than just discuss them every few years. We hope that the presentation as a set of recommendations will provoke this discussion, leading to an adoption of the recommendations (or a version of them) to be the guiding principles by which we carry out optimization research.

2.2. Other disciplines

Other disciplines are not without their faults.

In the medical area Ioannidis (2005) considered 45 highly cited papers, finding that 7 (16%) were later contradicted and another 7 had reported effects that were stronger than those reported in later studies. Twenty (44%) have been replicated in later studies with the remaining 11 (24%) not having been challenged to date.

In forecasting, Evanschitzky and Armstrong (2010) studied the replications that were carried across two forecasting journals between 1996 and 2008. They found that just over a third (35.3%) of the replications confirmed the results of the initial study, 45.1% provided partial support and 19.6% provided no support. Given that about one-fifth of replication studies do not provide support for the original study, Evanschitzky and Armstrong suggest that there is a need for more replication studies. They also say that 'journals can aid this by requiring and archiving full disclosure details and by inviting authors to replicate specific papers'.

A study going back to 1994 (Hubbard and Armstrong, 1994) showed that of 1120 papers sampled none were replications of previous work. Twenty of the papers were extensions, with 12 of those conflicting with the original work and only 3 confirming the previous work. This study was extended in 2007 (Evanschitzky *et al*, 2007), with the authors noting that the editorial policies of some journals now encourage replication. However, they note: 'Results show that the replication rate has fallen to 1.2%, a decrease in the rate by half. As things now stand, practitioners should be skeptical about using the results published in marketing journals as hardly any of them have been successfully replicated'.

In the disciplines of Accounting, Economics, Finance, Management and Marketing, Hubbard and Vetter (1996) carried out an an analysis of 18 business journals, covering 1970–1996. It showed that replication studies are not common in the business disciplines but when they are carried out the results are often at odds with the original study.

Easley *et al* (2000) say that 'The role of replication in marketing research has been a tenuous one at best'. They argue that the absence of replication research in the social sciences is the result of incorrect perceptions about how the studies should be carried out and how acceptable these studies would be.

Nature has recently published a series of articles around the idea of *Reducing our irreproducibility* (Anon, 2013).[1] One of the articles focuses on the importance of reproducibility (Russell, 2013), arguing that grant funding should be tied to be able to reproduce results, presenting a series of counter arguments against what would be seen as an unpopular proposal by some in the community. Vaux (2012) argues that many papers contain incorrect or sloppy statistics, which leads to sloppy science. The paper also serves as a good point of reference for a general introduction to statistics and we note that there are some excellent papers in the *Operations Research* community which outline various statistical tests and how they can be used (eg, Taillard *et al*, 2008).

Begley and Ellis (2012) argue that standards need to be raised in pre-clinical trials for cancer research. They report that only 25% of published pre-clinical trials could be validated to the point at which projects could continue. Perhaps, more worrying, is that the irreproducible papers had spawned many other papers that were based on the original findings. In order to counter such issues, Baker (2012) says that some scientific publishers are supporting an initiative, which would request high profile authors to have the results verified by an independent laboratory. Prinz *et al* (2011) say that many pharmaceutical

[1]The full series of articles are freely available at go.nature.com/huhbyr

companies run in house validation programmes before making a significant investment. However, they often find that exciting results in a scientific paper are difficult to reproduce. The paper concludes that literature data on potential drugs should be viewed with caution and the importance of carrying out a verification test.

2.3. Good laboratory practice

An internationally recognized definition of GLP is available on the The Medicines and Healthcare Products Regulatory Agency (MHRA) website. This organization carries the responsibility for regulating all medicines and medical devices in the United Kingdom.

> *'Good Laboratory Practice (GLP) embodies a set of principles that provides a framework within which studies are planned, performed, monitored, recorded, reported and archived. These studies are undertaken to generate data by which the hazards and risks to users, consumers and third parties, including the environment, can be assessed for pharmaceuticals, agrochemicals, veterinary medicines, industrial chemicals, cosmetics, food and feed additives and biocides. GLP helps assure regulatory authorities that the data submitted are a true reflection of the results obtained during the study and can therefore be relied upon when making risk/safety assessments.'*[2]

GLP was established following four scientists being put on trial after faking drugs and chemical studies during the 1970s (Marshall, 1983). GLP was introduced in 1972 in New Zealand and Denmark and then in 1978 in the United States. The principles were later adopted by the Organization for Economic Co-operation and Development (OECD) in 1992. The OECD has helped to promote the principles in many countries since that time.

3. Good Laboratory Practice for Optimization Research —GLP4OPT

In this section we present the recommendations that should be adopted by the optimization research community. Many of these have been suggested in other works, including many of those referred to in Section 3.1. However, to our knowledge, there has been very little take up on many of the proposals made and many of the bad practices that were highlighted still exist. We hope that providing an explicit set of recommendations will focus the minds of researchers as to the areas that need to be addressed when proposing a hypothesis and then designing and running simulations to investigate that hypothesis. The recommendations have been split into a number of key areas.

[2]http://www.mhra.gov.uk/Howweregulate/Medicines/Inspectionandstandards/GoodLaboratoryPractice/Structure/index.htm, last accessed 5 November 2013.

3.1. Guiding principles

There are certain guiding principles that underpin all the other recommendations. This is similar to the guidelines set out by the Committee on Publication Ethics (COPE) (http://publicationethics.org/), which is aimed at journal editors and publishers.

In some sense, they are *obvious*, but we believe that it is worth stating them explicitly, so that there is no confusion.

Recommendations for GLP4OPT

Recommentation 1 Researchers should adopt the highest ethical standards in conducting their research.

Recommentation 2 Researchers should ensure that the work of others is properly cited.

3.2. Benchmark data sets

Some of the papers we reviewed in Section 3.1 were critical of the way that benchmark instances were either introduced into the scientific literature, were biased towards certain algorithms or that randomly generated instances were not a true representation of the problem being considered or that the randomness was biased in some way. Moreover, instances are not always available for other researchers to use and, when randomly generated instances were used the algorithm is not reproducible, meaning that other researchers cannot generate a representative set of instances. Indeed, any generated instance, or its generator, should either be made available to other researchers or the algorithm (including any random number generation) should be defined to enable another researcher to create the same instance.

There are many data sets that have been introduced into the scientific literature. There should be some control as to how a data set is introduced into the scientific literature, and ultimately utilized.

Recommendations for GLP4OPT

Recommendation 3 Where possible, data should come from the real world, for example, from biological experiments.

Recommendation 4 Any benchmark data set that is accepted into the scientific community should be stored in one central location (see Section 4.16) so that all the data sets are available to all researchers in one place. Associated with each instance will be the paper (see Recommendation 5) that introduced the benchmark, along with all published results.

Recommendation 5 Any data set that a researcher wants to become a recognized data set should be subject to the same peer review process as any other scientific paper. Furthermore, where possible, a data set should be introduced via a stand-alone paper so that the peer review process can focus on the proposed benchmark instances, rather than on an algorithm that is being proposed. The paper should introduce the data sets, and provide all the elements that are set out in Recommendations 6–14.

Recommendation 6 A benchmark instance should be presented in a standard format, that is specified in such a way that new instances can be presented in the same way. We recognize that this may not be possible for existing data sets, as they may have different formats, but when a new data set is introduced, the authors should define how future instances should be represented.

Recommendation 7 The format in which solutions should be presented should be defined. The format should be amenable to parsing by a computer. We recognize that this may not be possible for existing data sets, as they may have different formats, but when a new data set is introduced, the authors should define how future instances should be represented.

Recommendation 8 When a benchmark instance is proposed, the researcher should provide the necessary means to evaluate each instance. As a minimum a class should be provided in as many languages as possible (eg, Java, C++, C#, PHP, Python) which the researcher can utilize to validate the validity and solution quality of any instance that they create. Several examples should be given, to show the evaluation of a number of solutions to help researchers verify that they are evaluating the solutions they produce in the correct way (see Recommendation18).

Recommendation 9 When a benchmark instance is proposed, a statistical test should also be suggested that can be used by future researchers to compare against other instances from the benchmark set. This should provide a way to compare different algorithms in a robust way. There are many sources of reference for defining suitable statistical tests. Recent examples include Derrac *et al* (2011) and García *et al* (2010).

Recommendation 10 When a benchmark instance is proposed, the researcher should state the recommended computational times used for each instance. Two main timing measures could be adopted. First, the number of times that the evaluation function can be called. Second, how long the algorithm can run using the same time as returned by Recommendation 28.

Recommendation 11 Researchers are encouraged to present a range of different benchmark instances, covering different sizes and complexity. Some of the instances may be solvable to optimality (eg, using mathematical programming) so that stochastic methods are able to provide some indication how close they are to optimality. Other instances might be challenging to the methodologies available today and others might be considered to provide challenges for the foreseeable future.

Recommendation 12 At the same location (see Section 4.16), where the benchmark instances are located, a history of best known solutions should be available. This, by definition, means that the best known solution is always available.

Recommendation 13 Any paper that is *published* (not just under review) that utilizes one of the standard benchmark data sets must provide ALL the solutions that are referred to in the paper that the authors have published. As an example, if the paper produces 30 solutions for statistical analysis then all those 30 solutions should be available so that other researchers are able to perform comparisons against the set of solutions rather than just the best, mean and median (also see Recommendation 41). Note, we are not suggesting that all the instances in a given data set have to be reported. A researcher might decide not to use all instances. But, if any instance is used, then ALL the solutions should be made available to the community.

Recommendation 14 It is recognized that there are some data sets that have been in the scientific literature for so long that they have become a *de facto standard*. We cannot simply ignore these and the international advisory board (see Section 6) will define those data sets that should form the basis for this proposed initiative.

3.3. Non-registered data sets

Not every researcher will want to register a benchmark data set, as defined in Section 4.2. For example, they may not want to wait for a separate paper to be peer reviewed enabling them to claim that they are using a registered data set. Indeed, they may not be convinced that their data set will stand up to the scientific review process but they still believe that the data set is important in the context of their work.

Recommendations for GLP4OPT

Recommendation 15 If a researcher decides not to register a data set that they use, they should still make the data set available to the scientific community and should seek to respect Recommendations 6–13.

3.4. Presentation of solutions

Too often in the optimization research literature, when presenting the results of a simulation the researcher simply reports the value of the evaluation function. If another researcher wants to view the solution, to either verify that the evaluation is correct, or to simply inspect it to give insights for further research, the solution is often impossible to access. Occasionally, results are verified. A good example of this is the Traveling Tournament Problem (TTP) instances, introduced in Easton *et al* (2001). Before a result can be recognized on the web site (http://mat.gsia.cmu.edu/TOURN/), the solution has to be sent to Michael Trick, one of the authors who established the benchmarks in Easton *et al* (2001), who will validate that it is correct.

Recommendations for GLP4OPT

Recommendation 16 When presenting the result of a simulation, any solution that is presented *must* be accompanied by the actual solution in the format that can be validated by

Recommendation 8. The solutions that are presented (both the solution and its evaluation) should include all those referred to in the paper. For example, any solutions that are used in statistical analysis, or when tuning parameters (see Recommendation 41).

3.5. Evidence of hybridized approaches

Many researchers are now using hybridized approaches to develop effective algorithms. It is important to provide evidence that the hybridized approach is actually effective.

Recommendations for GLP4OPT

Recommendation 17 If a hybridized algorithm is being presented then, unless previously reported in other work, the researchers should provide results from each *non-hybridized* element, using statistical analysis so that there is evidence that the hybridization is actually effective.

3.6. Evaluation function

The evaluation function is often the most important part of the paper and it is important to get this aspect of the paper correct.

Recommendations for GLP4OPT

Recommendation 18 The evaluation function must be described in a way that can be understood and implemented by other researchers.

Recommendation 19 Referring to Recommendation 8 it should be possible for another researcher to test their implementation of an evaluation function to check that they have implemented it correctly and that it produces the same values as Recommendation 8.

3.7. Algorithm presentation

It is important that an algorithm is presented in a way that another researcher could reproduce it.

Recommendations for GLP4OPT

Recommendation 20 Any algorithm that is presented must be reproducible. This is one of the underpinning tenets of scientific research but some details are often missing, or vague. The optimization research community needs to ensure that all algorithms (and results) are reproducible.

Recommendation 21 An algorithm should be presented in pseudo code, rather than a specific programming language. This is to stop any assumptions being made about another researchers' understanding on any given programming language.

Recommendation 22 Attention should be given to areas such as variable initialization, how many iterations of certain lines are carried out, all (sub-)operations are defined and so on.

3.8. Computational times

Recommendations for GLP4OPT

Computational times have been an issue for a long time, and will remain so for the foreseeable future. The recommendations in this section, we hope, will resolve many of the issues.

Recommendation 23 The researcher should adopt the recommended computational times as provided in Recommendation 10, stating which method they are using.

Recommendation 24 A researcher is at liberty to use a different computational time, but it must be done in such a way that other researchers are able to use the same timings in future research (see Recommendation 28)

3.9. Statistics

The question of which statistics to use is often problematical. However, if Recommendation 9 is adopted then each registered benchmark will come with a set of recommendations as to how different algorithms can be compared. Moreover, this recommendation would have undergone a peer review process so the community will have acknowledged that this statistical test is sufficient. In addition, Recommendation 16 will ensure that there is a sufficient number of solutions available (rather than just the best solution) so that the relevant statistical test has the correct data in order to be applied.

Recommendations for GLP4OPT

Recommendation 25 The necessary statistical test, as defined by Recommendation 9, should be applied to the relevant results (eg, the best known solution and the solutions produced from the current study).

3.10. Experimental conditions

One of the issues in comparing different algorithms is the environment on which those simulations were run. Whilst we can never account for different programming styles/skills there is a lot we can do to, at least, record the differences in the environment that other researchers will use in the future.

Recommendations for GLP4OPT

Recommendation 26 The full specification of the machine that was used for the experiments should be given. This should include the type of computer, the operating system and version number, the programming language that was used together with the compiler, the amount of memory available to the programme, whether parallel processing was employed, if GPUs were utilized and so on. Increasingly, researchers are using cloud services. These should also be reported in appropriate detail.

Recommendation 27 The date(s) that the experiments were run should be given. This provides another point of reference for future researchers. This is important as the

time between running experiments and the paper appearing in print could be a matter of years.

Recommendation 28 For each main operating system the community will provide a test programme (Dongarra, 1992, might provide some inspiration for how this can be done) that can be run on a researcher's target machine. This will record the speed of the researcher's machine against some known baseline, informing the researcher how long they should run their algorithm to give a fair comparison with other researchers, particularly the original benchmark. See McCollum *et al* (2010) for an example of this method being used for an international competition.

The benchmark programme will also provide a data file that can be stored as part of the paper that is subsequently published which future researchers can refer to in order to compare their algorithm's performance against previous algorithms.

3.11. Laboratory notebooks

As is the case with other scientific disciplines, optimization researchers should be encouraged to maintain a laboratory notebook and make these available to interested parties. Like some of the other recommendations, this is good scientific practice but is noted here so that there is no confusion.

Recommendation 29 Optimization researchers should maintain a laboratory notebook. These can be used by the researcher to record their thoughts, ideas and successes/failures and they also provide an important historic record for future researchers to refer to.

Recommendation 30 Researchers are encouraged to make available, via a supplementary file, the relevant parts of their notebooks, once a paper has been published.

Recommendation 31 Other researchers should not be critical of notebooks, but should accept them for what they are. A notepad that provides a historic document of the research ideas, thoughts and timeline associated with a given piece of research.

3.12. Software

Whether software should be provided as part of the peer review process, and subsequently made available to other researchers, is a moot point in optimization research, although we note that Bioinformatics generally insists on software being made available. There have been recent suggestions that software should be provided. Ince *et al* (2012) argue that 'anything less than the release of source programmes is intolerable for results that depend on computation'.

Having access to the software of another researcher would save a lot of time as algorithms would not have to be reimplemented. Moreover, it would remove one area where mistakes could be introduced due to errors/misunderstandings in implementation or the algorithm not being explained well enough to ensure that it can be reproduced.

On the other hand, many researchers are reluctant to provide their code as they might be embarrassed about their coding skills/style and they may feel that they might be asked to explain certain parts of the code or, indeed, provide support.

The fact that an algorithm works now, might not be the case when a new environment is used (eg, new operating systems, different compiler versions, etc). There may also be commercially sensitive reasons why the code should not be in the public domain and researchers may not want to make the code available which has been the result of an investment of weeks/months/years in developing the software.

However, we believe that software which underpins a scientific paper should be part of the peer reviewed scientific archive. This not only adds to the integrity of the discipline but also provides a historical reference that can be utilized by future researchers, perhaps for reasons other than algorithm reproducibility. This might include an historical perspective of programming languages, how coding styles have changed and how the implementations of the same algorithms differ.

Recommendations for GLP4OPT

Recommendation 32 It is beholden on the individual researcher(s) to keep a copy of the software that led to the contribution claimed in a given paper. Researcher(s) should retain an *exact* copy of the source code, the compiled code and any other environmental factors that are required to reproduce their results. If necessary, they should be able to re-run their algorithm and produce *exactly* the same results (see Recommendation 33), even years later.

Recommendation 33 In the case of any dispute, authors can be asked to demonstrate their software to show that it produces the results claimed in the paper. This request can only come from the editor-in-chief (or their nominated representative) of the journal to which a paper has been submitted to, or published in.

Recommendation 34 The source code for a given paper should be made available, under one of four categories:

(1) The source code is made freely available to other researchers.
(2) The source code is not available to other researchers, but is available to the reviewers who can access the code using the accepted confidentially that is associated with reviewing scientific papers.
(3) The source code is supplied, but is not made available to either researchers or the reviewers.
(4) The source code is not made available at all.

Unless the first option is chosen, a statement/justification must be given in the paper as to why the source code is being restricted. This statement will be subject to review

and may be reason to recommend that the paper not be published.

Recommendation 35 Supplying source code does imply that any support will be given to other researchers who later use, or access, the source code.

Recommendation 36 The complete environment must be described so that anybody using the source code is able to replicate the computational conditions. As an example, the experimental setup might utilize Matlab, R or mathematical software. These elements need to be described in a manner that enables reproducibility.

Recommendation 37 Source code associated with a published paper should be held in a repository that does not allow updates (any corrections should be done via an erratum) and which can be accessed via a *permalink*, in the same way that scientific papers are generally available via a DOI. This is to ensure that, like scientific papers, scientific researchers are accessing the same version of the source code as supplied by the authors.
In order to protect the scientific archive, the repository (or repositories) should be maintained by the scientific community, rather than an Open Source repository.

Recommendation 38 Only in exceptional circumstances will access be given to the code without the author's permission, and then only the editor-in-chief of the relevant journal can authorize this. Exceptional circumstances might include accusations of plagiarism, or suspicions of incorrect results being claimed.

Recommendation 39 If all the authors have passed away, the software will be available to anybody that requests it, unless other arrangements have been made. This is to allow future researchers to access the code without any hindrance. There is no expectation of support of any kind from colleagues, or institutions.

3.13. Parameter selection and values

All too often, researchers state that parameter values were set by *initial experimentation*, arriving at the values used in the paper. This misses many points of interest to the scientific community. For example:

- We are often not told what experiments were carried out to arrive at the values that were used.
- We are often not informed how many experiments were run, or how long they took to arrive at the final values.
- There is often no statistical analysis, comparing one set of parameters against another in order to demonstrate that a given set of parameter values is *better* than another or indeed whether the parameters are insensitive to change which could indicate a robust algorithm.
- There is often no comparison with samples from outside the tested data set to provide evidence, or not, that the parameters are robust to problem instances outside of those that are the real focus of the study.

Recommendations for GLP4OPT

Recommendation 40 Be specific on how the parameter values were set, detailing the simulations that were carried out, to the point that these should be as reproducible as any other simulations reported in the paper.

Recommendation 41 All the solutions that are used to compare the parameters should be made available to other researchers in the same way that the main solutions are also reported (see Recommendation 16).

Recommendation 42 When making claims that one set of parameter values is better than another, a statistical analysis should support this claim.

Recommendation 43 If no evidence is provided to the contrary, no claims should be made that the parameter values that were used for the main simulations presented in the paper are in any way robust, optimal or can be relied upon in other simulations to provide any level of solution quality.

3.14. Random number generation

Many papers simply state that a random number is generated, being drawn from a given distribution (eg, normal, uniform, etc). Not only is the random number generator (RNG) key to reproducibility but the RNG that is used may have certain (sometimes undesirable) properties (Hellekalek, 1998). It should be noted that random number generation is applicable to both search methodologies, as well as instance generation.

Recommendations for GLP4OPT

Recommendation 44 The RNG that was used should be explicitly stated. The coding lines that are used to generate each random variable should be presented, along with any preparatory steps (eg, setting starting values, seeding the sequence).

Recommendation 45 It should be possible for any researcher to generate the same sequence of random numbers, so that the simulation can be reproduced.

3.15. Termination criteria

In many papers it is often stated to carry out a sequence of algorithmic steps until some termination criteria are met. The exact criteria are often obscure or several choices are given and it is not clear which one has been used. For example, 'repeat steps n to m until a certain period of time has passed or q iterations have been performed'. It should be possible to

precisely ascertain the termination criteria that were used for any simulation that is presented.

Recommendations for GLP4OPT

Recommendation 46 Researchers should precisely define the termination criteria for any algorithm that is presented.

3.16. GLP4OPT adherence and support

Evidence to demonstrate compliance with GLP4OPT should be provided as a supplementary file(s) to the main article. This enables authors to show compliance, without having to take up valuable space in the main article. In line with current practice, any supplementary files are subject to peer review.

If the recommendations in this paper are embraced by the community, this might lead to a web presence for GLP4OPT, which could reference, and link to, all the papers that have been peer reviewed in-line with the recommendations contained in this paper.

There may also be an opportunity for one of the *Operations Research* societies to lead on the GLP4OPT standard but we recognise that this is not a decision that is within the scope of this paper and would be subject to much more discussion.

In the remainder of this section, we provide specific recommendations as to what should constitute a supplementary file(s) and how a website GLP4OPT might support the community in the future. Section 6 provides more thoughts as to how GLP4OPT might be progressed.

Recommendations for GLP4OPT

Recommendation 47 Any benchmarks that are introduced (see Recommendations 5–14) into the scientific literature must be recorded in the journal's supplementary repository (see Recommendation 54).

Recommendation 48 The GLP4OPT supplementary file must define which version of the standard has been used. As mentioned in the Introduction, the version introduced here is *GLP4OPT version 1.00*.

Recommendation 49 If a journal or publisher has subscribed to GLP4OPT then it should insist on a supplementary file that reviewers can access to check against compliance with the standard. It should also be made available to those readers that are enable to access the main paper.

Recommendation 50 Once (and if) established, the GLP4OPT website will maintain the best known solutions for each registered data set, along with a history of best known solutions (see Recommendation 12).

Recommendation 51 ALL solutions that are referred to in a given paper must be uploaded to the journal as a supplementary file(s) (see Recommendation 13). This includes any solutions used to establish parameter settings (see Recommendation 41)

Recommendation 52 Once (and if) established, the GLP4OPT website will contain a repository of all the papers that have used a given registered benchmark data set.

Recommendation 53 Once (and if) established, the GLP4OPT website will maintain the set of historical data sets (see Recommendation 14), including best known solutions and all papers that have used those data sets.

Recommendation 54 It is recognized that many of the data specified in this section might be held as a supplementary file on the journal's website. In this case, the GLP4OPT website will contain the necessary links to the supplementary data.

4. Reviewing against GLP4OPT

If the GLP4OPT standard was adopted, it would lead naturally to a set of prescribed questions that any reviewer might ask, and which authors might be expected to respond to.

Of course, a negative answer to any of these questions would not automatically disqualify the paper from being published. That decision still ultimately lies with the reviewers and the editors, but it might help provide additional focus/information for the review process. Moreover, knowing that the reviewers will explicitly look at these aspects of the paper may help the authors in planning their experiments and the reporting of their work.

5. Progression

In this section, we outline several ideas as to how the views expressed in this paper could be progressed in order to provide the optimization research community with a more robust experimental environment.

In our view, it is the lack of a progression plan that has hindered anything happening until now. The ideas that have been expressed since the 1970s have all been very valid proposals but it was left to the community to adopt them. This has not happened, or only to a very limited extent.

To progress the adoption of the proposals outlined in this paper, we suggest the following, once this paper has been published.

(1) Publishers and journals are approached, asking them to adopt the standards and requesting that they incorporate the standard into the work flow, as well as informing authors, reviewers and editors that the standard has been adopted. This would be similar to them adopting COPE (Committee on Publication Ethics).
(2) *Operations Research* societies are approached asking that they adopt this standard, and promote its use to it members.
(3) Funding agencies are approached to ask if they would be willing to adopt GLP4OPT, specifying that their use is a condition of any grant that they award.

(4) The views/opinions of the entire community (including the industrial community) should be sought to ask their ideas/feedback on the suggested GLP4OPT standards. This will lead to a new version of GLP4OPT.
(5) *Operations Research* societies are approached to ask if one of them would like to take the lead on maintaining, developing and promoting GLP4OPT.
(6) A Management Board could be established that oversees GLP4OPT, in the same way that OECD acts for GLP for non-clinical research. The Management Board will be responsible for (but not limited to) raising funds, maintaining and developing GLP4OPT, liaising with the community (authors, readers, publishers, societies, etc) and ensuring that GLP4OPT is adopted.
    The Management Board should be international in make up, should include experienced and early career researchers, as well as the industrial community. The Management Board must also represent all of optimization research.
(7) An International Advisory Board could be established. The make up, and remit, of this board will be the responsibility of GLP4OPT's Management Board but it should certainly be international in nature and should provide coverage across all of optimization research, and have representation from experienced and early career researchers as well as the industrial community.
(8) A website could be established to support GLP4OPT.

None of the authors of this paper are suggesting that, as authors of the paper, that they should serve in any particular role with regard to GLP4OPT. If this paper is published, the authors will progress GLP4OPT until suitable structures have been established.

6. Conclusions

This paper provides the opportunity for those that embrace the ideas outlined in this paper (or those that are developed as a result of this paper) to work to a common set of standards which are recognized across the disciplines. Even though they may not be perfect, and subject to future enhancements, at least we will all be working to the same set of imperfect standards, rather than everybody working to a different set.

We realize that papers such as these are not without controversy and that there will never be a set of guiding principles that everybody agrees to and, of course, any research group or individual has the right to simply ignore any recommendations that arise as a result of this paper. But, we hope that, if any of the principles outlined in this paper are adopted, the community would recognize the potential benefit both in terms of carrying out and reporting its research but also providing training to its researchers in carrying out high quality research.

We also hope that adherence to the GLP4OPT standards, would give other researchers some level of confidence as the quality of the research that has been conducted but it would also provide future researchers with a set of data that will enable them to more easily compare against their own research. Moreover, it would build up a valuable set of data that could be used in a variety of ways in the future, above and beyond, simply using the benchmark data to compare against a current research idea.

References

Adenso-Diaz B and Laguna M (2006). Fine-tuning of algorithms using fractional experimental designs and local search. *Operations Research* **54**(1): 99–114.

Ahuja RV and Orlin J (1996). Use of representative operation counts in computational testing of algorithms. *INFORMS Journal on Computing* **8**(3): 318–330.

Anon (2013). Announcement: Reducing our irreproducibility. *Nature* **496**(7446): 398.

Archetti C and Speranza MG (2008). The split delivery vehicle routing problem: A survey. In: Golden B, Raghavan S and Wasil E (eds). *Vehicle Routing Problem: Latest Advances and New Challanges, Operations Research Computer Science Interfaces*. Vol. 43, Springer, New York, pp 103–122.

Baker M (2012). Independent labs to verify high-profile papers: *Nature News*, 14 August 2012, Chicago. doi:10.1038/nature.2012.11176.

Balinski ML (1978). On the reporting of computational experiments. *Mathematical Programming* **15**(1): 315.

Barr RS, Golden BL, Kelly JP, Resende MGC and Stewart Jr WR (1995). Designing and reporting on computational experiments with heuristic methods. *Journal of Heuristics* **1**(1): 9–32.

Begley CG and Ellis LM (2012). Drug development: Raise standards for preclinical cancer research. *Nature* **483**(7391): 531–533.

Bellmore M and Nemhauser GL (1968). Travelling salesman problem: A survey. *Operations Research* **16**(3): 538–558.

Błazewicz J, Domschke W and Pesch E (1996). The job shop scheduling problem: Conventional and new solution techniques. *European Journal of Operational Research* **93**(1): 1–33.

Błazewicz J, Ecker KH, Pesch E, Schmidt G and Weglarz J (2007). *Handbook on Scheduling*. Springer–Verlag: Berlin, Heidelberg.

Błazewicz J, Formanowicz P and Kasprzak M (2005). Selected combinatorial problems of computational biology. *European Journal of Operational Research* **161**(3): 585–597.

Boylan JE, Goodwin P, Mohammadipour M and Syntetos AA (2015). Reproducibility in forecasting research. *International Journal of Forecasting* **31**(1): 79–90.

Braysy O, Dullaert W and Gendreau M (2004). Evolutionary algorithms for the vehicle routing problem with time windows. *Journal of Heuristics* **10**(6): 587–611.

Braysy O and Gendreau M (2005a). Vehicle routing problem with time windows, part I: Route construction and local search algorithms. *Transportation Science* **39**(1): 104–118.

Braysy O and Gendreau M (2005b). Vehicle routing problem with time windows, part II: Metaheuristics. *Transportation Science* **39**(1): 119–139.

Brown L (1999). *Technical and Military Imperatives: A Radar History of World War 2*. CRC Press: Bristol, UK.

Brucker P, Sotskov YN and Werner F (2007). Complexity of shop-scheduling problems with fixed number of jobs: A survey. *Mathematical Methods of Operations Research* **65**(3): 461–481.

Burkard RE, Deineko VG, Van Dal R, Van der Veen JAA and Woeginger GJ (1998). Well-solvable special cases of the traveling salesman problem: A survey. *SIAM Review* **40**(3): 496–546.

Burke EK *et al* (2013). Hyper-heuristics: A survey of the state of the art. *Journal of the Operational Research Society* **64**(12): 1695–1724.

Cattrysse DG and Van Wassenhove LN (1992). A survey of algorithms for the generalized assignment problem. *European Journal of Operational Research* **60**(3): 260–272.

Cheng RW, Gen M and Tsujimura Y (1999a). A tutorial survey of jobshop scheduling problems using genetic algorithms, part II: Hybrid genetic search strategies. *Computers & Industrial Engineering* **36**(2): 343–364.

Cheng RW, Gen M and Tsujimura Y (1999b). A tutorial survey of jobshop scheduling problems using genetic algorithms: Part II. Hybrid genetic search strategies. *Computers & Industrial Engineering* **37**(1–2): 51–55.

Cook WJ (2011). *In Pursuit of the Traveling Salesman: Mathematics at the Limits of Computation*. Princeton University Press: USA.

Cordeau JF, Gendreau M, Laporte G, Potvin JY and Semet F (2002). A guide to vehicle routing heuristics. *Journal of the Operational Research Society* **53**(5): 512–522.

Coy SP, Golden BL, Runger GC and Wasil EA (2001). Using experimental design to find effective parameter settings for heuristics. *Journal of Heuristics* **7**(1): 77–97.

Crowder H, Dembo RS and Mulvey JM (1979). On reporting computational experiments with mathematical software. *ACM Transactions on Mathematical Software* **5**(2): 193–203.

Crowder HP, Dembo RS and Mulvey JM (1978). Reporting computational experiments in mathematical programming. *Mathematical Programming* **15**(1): 316–329.

Derrac J, García S, Molina D and Herrera F (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation* **1**(1): 3–18.

Doerner KF and Schmid V (2010). Survey: Matheuristics for rich vehicle routing problems. In: Blesa MJ, Blum C, Raidl G, Roli A and Sampels M (eds). *Hybrid Metaheuristics, Lecture Notes in Computer Science*. Vol. 6373, 7th International Workshop on Hybrid Metaheuristics, Vienna, Austria, Springer: Berlin Heidelberg, pp 206–221, Oct 01-02, 2010.

Dongarra JJ (1992). Performance of various computers using standard linear equations software. *ACM SIGARCH Computer Architecture News* **20**(3): 22–44.

Easley RW, Madden CS and Dunn MG (2000). Conducting marketing science: The role of replication in the research process. *Journal of Business Research* **48**(1): 83–92.

Easton K, Nemhauser G and Trick M (2001). The traveling tournament problem description and benchmarks, chap. Principles and Practice of Constraint Programming CP 2001: *7th International Conference*, CP 2001 Paphos, Cyprus, 26 November–1 December, 2001 Proceedings. Lecture Notes in Computer Science 2239, Springer, Berlin Heidelberg, pp 580–584.

Ernst AT, Jiang H, Krishnamoorthy M and Sier D (2004). Staff scheduling and rostering: A review of applications, methods and models. *European Journal of Operational Research* **153**(1): 3–27.

Evanschitzky H and Armstrong JS (2010). Replications of forecasting research. *International Journal of Forecasting* **26**(1): 4–8.

Evanschitzky H, Baumgarth C, Hubbard R and Armstrong JS (2007). Replication research's disturbing trend. *Journal of Business Research* **60**(4): 411–415.

Fildes R (1979). Quantitative forecasting the state of the art: Extrapolative models. *Journal of the Operational Research Society* **30**(8): 691–710.

Fildes R (1985). Quantitative forecasting the state of the art: Econometric models. *Journal of the Operational Research Society* **30**(7): 549–580.

Fildes R, Nikolopoulos K, Crone SF and Syntetos AA (2008). Forecasting and operational research: A review. *Journal of the Operational Research Society* **59**(9): 1150–1172.

García S, Fernández A, Luengo J and Herrera F (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences* **180**(10): 2044–2064.

Gass SI and Assad AA (2006). *An Annotated Timeline of Operations Research: An Informal History*. Springer: New York.

Gendreau M, Potvin J-Y, Braysy O, Hasle G and Lokketangen A (2008). Metaheuristics for the vehicle routing problem and its extensions: A categorized bibliography. In: Golden B, Raghavan S and Wasil E (eds). *Vehicle Routing Problem: Latest Advances and New Challenges, Operations Research Computer Science Interfaces*. Vol. 43, Springer, New York, pp 143–169.

Ghobbar AA and Friend CH (2003). Evaluation of forecasting methods for intermittent parts demand in the field of aviation: A predictive model. *Computers & Operations Research* **30**(14): 2097–2114.

Golden BL, Assad AA, Wasil WA and Baker E (1986). Experimentation in optimization. *European Journal of Operational Research* **27**(1): 1–16.

Golden BL and Stewart WR (1985). *The Traveling Salesman Problem*. chap. Empirical Analysis of Heuristics, John Wiley & Sons: Chichester, UK, pp 207–249.

Greenberg HJ (1990). Computational testing: Why, how and how much. *ORSA Journal on Computing* **2**(1): 94–97.

Hellekalek P (1998). Good random number generators are (not so) easy to find. *Mathematics and Computers in Simulation* **46**(5–6): 485–505.

Hoffman A, Mannos M, Sokolowsky D and Wiegmann N (1953). Computational experience in solving linear programs. *Journal of the Society for Industrial and Applied Mathematics* **1**(1): 17–33.

Hooker JN (1994). Needed: An empirical science of algorithms. *Operations Research* **42**(2): 201–212.

Hooker JN (1995). Testing heuristics—We have it all wrong. *Journal of Heuristics* **1**(1): 33–42.

Hooker JN (2007). Good and bad futures for constraint programming (and operations research). *Constraint Programming Letters* **1**: 21–32.

Hubbard R and Armstrong JS (1994). Replications and extensions in marketing: Rarely published but quite contrary. *International Journal of Research in Marketing* **11**(3): 233–248.

Hubbard R and Vetter DE (1996). An empirical comparison of published replication research in accounting, economics, finance, management, and marketing. *Journal of Business Research* **35**(2): 153–164.

Ignizio JP (1971). On the establishment of standards for comparing algorithm performance. *Interfaces* **2**(1): 8–11.

Ince DC, Hatton L and Graham-Cumming J (2012). The case for open computer programs. *Nature* **482**(7386): 485–488.

Ioannidis JA (2005). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association* **294**(2): 218–228.

Jackson RHF, Boggs PT, Nash SG and Powell S (1991). Guidelines for reporting results of computational experiments. Report of the *ad hoc* committee. *Mathematical Programming* **49**(1–3): 413–425.

Jackson RHF and Mulvey JM (1978). A critical review of methods for comparing mathematical programming algorithms and software (1953–1977). *Journal of Research of the National Bureau of Standards* **83**(6): 563–584.

Jaillet P and Wagner MR (2008). Online vehicle routing problems: A survey. In: Golden B, Raghavan S and Wasil E (eds). *Vehicle Routing Problem: Latest Advances and New Challenges, Operations Research Computer Science Interfaces*. Vol. 43, Springer, New York, pp 221–237.

Jourdan L, Basseur M and Talbi E-G (2009). Hybridizing exact methods and metaheuristics: A taxonomy. *European Journal of Operational Research* **199**(3): 620–629.

Kendall G, Knust S, Ribeiro CC and Urrutia SS (2010). Scheduling in sports: An annotated bibliography. *Computers & Operations Research* **37**(1): 1–19.

Kiran AS and Smith ML (1984). Simulation studies in job shop scheduling—1. A survey. *Computers & Industrial Engineering* **8**(2): 87–93.

Kirkpatrick S, Gelatt Jr CD and Vecchi M (1983). Optimization by simulated annealing. *Science* **220**(4598): 671–680.

Laporte G (2009). Fifty years of bvehicle routing. *Transportation Science* **43**(4): 408–416.

Lee C-Y, Bard J, Pinedo M and Wilhelm WE (1993). Guidelines for reporting computational resuts in IIE transactions. *IIE Transactions* **25**(6): 121–123.

Lin EYH (1998). A bibliographical survey on some well-known non-standard knapsack problems. *INFOR* **36**(4): 274–317.

Loiola EM, de Abreu NMM, Boaventura-Netto PO, Hahn P and Querido T (2007). A survey for the quadratic assignment problem. *European Journal of Operational Research* **176**(2): 657–690.

Lukasiak P, Błazewicz J and Milostan M (2010). Some operations research methods for analyzing protein sequences and structures. *Annals of Operations Research* **175**(1): 9–35.

Lust T and Teghem J (2010). The multiobjective traveling salesman problem: A survey and a new approach. In: Coello CAC, Dhaenens C and Jourdan L (eds). *Studies in Computational Intelligence*. Vol. 272, Springer-Verlag: Berlin, pp 119–141.

Lust T and Teghem J (2012). The multiobjective multidimensional knapsack problem: A survey and a new approach. *International Transactions in Operational Research* **19**(4): 495–520.

Marshall E (1983). The murky world of toxicity testing. *Science* **220**(4602): 1130–1132.

McCollum B *et al* (2010). Setting the research agenda in automated timetabling: The second international timetabling competition. *INFORMS Journal on Computing* **22**(1): 120–130.

McGeoch CC (1996). Toward an experimental method for algorithm simulation. *INFORMS Journal on Computing* **8**(1): 1–15.

Miller HE (1976). Personnel scheduling in public systems—A survey. *Socio-Economic Planning Sciences* **10**(6): 241–249.

Miller DM and Williams D (2003). Shrinkage estimators of time series seasonal factors and their effect on forecasting accuracy. *International Journal of Forecasting* **19**(4): 669–684.

Pentico DW (2007). Assignment problems: A golden anniversary survey. *European Journal of Operational Research* **176**(2): 774–793.

Pisinger D (2007). The quadratic knapsack problem—A survey. *Discrete Applied Mathematics* **155**(5): 623–648.

Prinz F, Schlange T and Asadullah K (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery* **10**(7391): 10–11.

Rardin RL and Uzsoy R (2001). Experimental evaluation of heuristic optimization algorithms: A tutorial. *Journal of Heuristics* **7**(3): 261–304.

Rasmussen RV and Trick MA (2008). Round robin scheduling—A survey. *European Journal of Operational Research* **188**(3): 617–636.

Russell JF (2013). If a job is worth doing, it is worth doing twice. *Nature* **496**(7443): 7.

Salkin HM and Kluyver CAD (1975). Knapsack problem—Survey. *Naval Research Logistics* **22**(1): 127–144.

Sörensen K (2015). Metaheuristics—The metaphor exposed. *International Transactions in Operational Research* **22**(1): 3–18.

Sörensen K and Glover F (2013). *Encyclopedia of Operations Research and Management Science*. chap. Metaheuristics, 3rd edn. Springer: New York.

Taillard ÉD, Waelti P and Zuber J (2008). Few statistical tests for proportions comparison. *European Journal of Operational Research* **185**(3): 1336–1350.

Talbi E-G (2009). *Metaheuristics: From Design to Implementation*. Wiley: Hoboken, NJ.

Trapero JR, Kourentzes N and Fildes R (2015). Identification of sales forecasting models. *Journal of the Operational Research Society* **66**(2): 299–307.

Vaux DL (2012). Research methods: Know when your numbers are significant. *Nature* **492**(7428): 180–181.

Vidal T, Crainic TG, Gendreau M and Prins C (2013). Heuristics for multiattribute vehicle routing problems: A survey and synthesis. *European Journal of Operational Research* **231**(1): 1–21.

Vines TH *et al* (2013). The availability of research data declines rapidly with article age. *Current Biology* **24**(1): 94–97.

Wilbaut C, Hanafi S and Salhi S (2008). A survey of effective heuristics and their application to a variety of knapsack problems. *IMA Journal of Management Mathematics* **19**(3): 227–244.

Zanakis SH, Evans JR and Vazacopoulos AA (1989). Heuristic methods and applications: A categorized survey. *European Journal of Operational Research* **43**(1): 88–110.

Received 27 October 2014;
accepted 19 August 2015 after one revision