Experimental Design & Methodology

Basic lessons in empiricism

Rafal Kicinger
rkicinge@gmu.edu

R. Paul Wiegand
paul@tesseract.org

EClab
George Mason University

Outline of Discussion

Part I:	Methodology	←
Part II:	Designing Experiments	
Example:	Exp. Methodology & Design	
Part III:	Conducting Experiments	
Part IV:	Presenting Experiments	
Example:	Conducting & Presenting Exp.	

A philosophy of research

- Research does not:
  - Consist of mere information gathering
  - Simply transport facts
  - Merely "rummage" for information
- Research *does*:
  - Originate with a question or problem
  - Require a clear articulation of a goal
  - Follow a specific plan or procedure (a *method*)
  - Require collection *and* interpretation of data
- Empirical research consists of:
  - Experimentation
  - Interpretation of results
  - Presentation of results

A philosophy of research

- Research does not:
  - Consist of mere information gathering
  - Simply transport facts
  - Merely ”rummage” for information
- Research *does*:
  - Originate with a question or problem
  - Require a clear articulation of a goal
  - Follow a specific plan or procedure (a *method*)
  - Require collection *and* interpretation of data
- Empirical research consists of:
  - Experimentation
  - Interpretation of results
  - Presentation of results
  - $\Rightarrow$ **Methodology**

Experimentation

- Why do we perform experiments?
  - [**Exploration**] Try to get our head around an issue
  - [**Comparison**] Compare two or more things (algorithms)
  - [**Explanation**] Explain how/why some property works
  - [**Demonstration**] Demonstrate a point, proof of concept, etc.
  - [**Theory Validation**] Validate some theoretical result
- For whom/what do we do so?
  - Ourselves
  - Publication

Experimentation

- Why do we perform experiments?
  - [**Exploration**] Try to get our head around an issue
  - [**Comparison**] Compare two or more things (algorithms)
  - [**Explanation**] Explain how/why some property works
  - [**Demonstration**] Demonstrate a point, proof of concept, etc.
  - [**Theory Validation**] Validate some theoretical result
- For whom/what do we do so?
  - Ourselves
  - Publication

Not the same motivation!

On Method

- What is method?
  - Clear, organized approach to scientific experimentation
  - Plan containing a source, goal, and path to get there
  - Collection of decisions about conducting experiments *and obtaining/interpreting results*
- Without (sound) method:
  - Restricted to mainly exploratory experimentation
  - Can gain intuition, but no real answers
  - Difficult to justify results to others
- With (sound) method:
  - Allow full range of types of experimentation
  - Can be used to determine clear answers
  - Facilitates justification of results

(Sound) Methodology

- Role of exploratory experimentation:
  - Only the initial, observational phase of experimentation
  - Not used to draw conclusions
  - May never appear in published materials
  - Used to help *generate* hypotheses
- Well-posed Questions
  - Questions should be clear, precise, and to the point
  - Questions should be tractable
  - Questions form the basis for hypotheses
  - Hypotheses should be *falsifiable*
  - Clear, justifiable results stem from experiments addressing a precise, well-posed question

(Sound) Methodology (2)

- Mechanistic details:
  - Clear statement of hypotheses
  - Experimental design
  - *A priori* decisions about result interpretation:
    - What are the assumptions and their potential ramifications?
    - What is being measured?
    - What is meant by qualitative terms (e.g., "better" or "best")?
    - How will outliers be removed?
    - What statistical tests will be run (why)?
    - What confidence levels will be used?
    - How many trials will be run?

(Sound) Methodology (2)

- Mechanistic details:
  - Clear statement of hypotheses
  - Experimental design
  - *A priori* decisions about result interpretation:
    - What are the assumptions and their potential ramifications?
    - What is being measured?
    - What is meant by qualitative terms (e.g., "better" or "best")?
    - How will outliers be removed?
    - What statistical tests will be run (why)?
    - What confidence levels will be used?
    - How many trials will be run?

Generally one should know (before the experiments are even run) what the possible outcomes are, and what those outcomes each mean in terms of the question.

Limits of Empiricism

- Empirical research (typically) cannot:
  - Answer a question not (or poorly) posed
  - Convince an audience of *fact*
  - Provide general answers

e.g., "Algorithm A *is always better than* B"

- Empirical research often *can*:
  - Answer a question clearly posed
  - Convince an audience of *probable fact*
  - Provide conditional answers

e.g., "Algorithm A *is usually better than* B *on problems with property* X"

Outline of Discussion

Part I:	Methodology	$\surd$
Part II:	Designing Experiments	$\longleftarrow$
Example:	Exp. Methodology & Design	
Part III:	Conducting Experiments	
Part IV:	Presenting Experiments	
Example:	Conducting & Presenting Exp.	

Selecting Problem Domain(s)

- Consider its relevance:
  - Does the question *center* around the problem domain?
  - What is the *point* of the problem domain?
  - What do you hope to learn?
  - What *cannot* be learned?
- Do not pick problems
  - Without reason or purpose
  - Just because it is in a common “Test Suite”
  - That are needlessly complicated, hard to understand
- Pick problems
  - That are simple, but salient
  - Demonstrative of particular property or properties
  - Illustrative of an “interesting” problem of study
  - Consistent with existing relevant studies
  - Analyzable, understandable, or (at least) intuitable

Selecting Algorithm(s)

- Consider its relevance:
  - Does the question center around (part of) the algorithm?
  - Does the question relate it to (properties of) the problem?
  - Are you comparing algorithms? What is the basis?
  - What can / cannot be learned?
- Do not pick algorithms
  - Without reason or purpose
  - Just because it is consistent with prior work *
  - That are needlessly complicated, hard to understand
- Pick algorithms
  - That are simple, but salient
  - That are consistent with prior work*
  - Demonstrating
    - Some quantifiable (or, at least, qualifiable) result
    - "Performance" under particular problem properties
    - A basis of comparison (apples to apples)
  - Analyzable, understandable, or (at least) intuitable

Constructing Experimental Groups

- Top-down design of groups
  - What are the “factors” of the experimental study?
  - What are the “levels” of these factors?
  - Develop a hierarchy based on problem and and algorithm?
  - Sketch out what you believe the results will be for groups if
    - Hypothesis is accepted
    - Hypothesis is rejected
- Important things to consider:
  - What is being compared?
  - Do you have control groups? What are they?
  - How much do “frivolous” groups cost you?
  - How important is turn-around time?
- Prioritize the groups
  - Prioritize by importance
  - Prioritize by turn-around need

Common EC Mistakes

- Problem domains (are) often
  - Very complicated in order to to be more "real-world"
  - Default to using De Jong test suite, without good reason
  - Use a vast number of problems to justify "generality"
- Algorithms (are) often
  - Poorly motivated (often unnecessarily complicated)
  - Excessively detailed in terms parameter values
  - Make naive choices for parameter values
  - Fail to compare against state of the art algorithms

Adjusting EA Parameters

- Sufficient for the task
  - Should be justifiable
  - Should be demonstrative of the point of study
  - When in doubt, use "traditional" settings
- Informal sensitivity studies
  - It is reasonable to do casual sensitivity studies to find "good" parameter values
  - Be careful to conclude nothing definitive from such a study
  - Watch for combinatorial explosion
    (you can't test everything)

Outline of Discussion

Part I:	Methodology	$\surd$
Part II:	Designing Experiments	$\surd$
Example:	Exp. Methodology & Design	$\leftarrow$
Part III:	Conducting Experiments	
Part IV:	Presenting Experiments	
Example:	Conducting & Presenting Exp.	

Epistasis in GAs

- Analysis of the role of epistasis in GAs: (Davidor, 1991)
- Type of research:
  - Explanatory
    - Determining the statistical properties of functions that make them suitable for GA optimization
    - Determining a degree of epistasis of a *given* problem

Epistasis

term used in genetics to denote the fact that the expression of a chromosome is not merely a linear function of the effects of its individual alleles.

Epistasis in GAs

- Research questions posed:
  - What properties of problems and their representations make them hard for GAs?
  - What is the influence of epistasis on the hardness of a problem?
  - How can we quantify the degree of epistasis for a *given* problem?
- Research goal:
  - Define (quantify) and explain the role of epistasis in GAs

Epistasis in GAs

- Davidor's Methodology:
  - Standard GA settings:
    - Binary representations
    - Fixed-length strings
    - Population of size N
  - Several statistical quantities defined:
    - Average fitness
    - Excess string fitness value
    - Average allele value
    - Excess allele value
    - Excess genic value
    - Genic value of a string
    - Epistatis measure

$$\bar{V} = \sum_{S \in Pop} v(S)/N$$

$$E(S) = v(S) - \bar{V}$$

$$A_i(a) = \sum v(S)/N_i(a)$$

$$E_i(a) = A_i(a) - \bar{V}$$

$$A(S) = \bar{V} + \sum E_i(a)$$

$$E(A) = \sum E_i(a)$$

$$\epsilon(S) = v(S) - A(S)$$

Epistasis in GAs

- Davidor's Methodology:
  - Estimating statistical quantities (variances):
    - Epistasis variance (for entire universe and population)
    - Fitness variance
    - Genic variance
  - Assumptions:
    - Information on many schemata can be processed in parallel
    - Schemata competitions can be isolated and solved independently
    - Combining small pieces of the genotype ('good' schemata) is a sensible method of finding optimal solutions
    - –> Schema Theorem

Epistasis in GAs

- ## Davidor's Methodology:
  - Hypotheses:
    - Epistasis for a given problem can be quantitatively measured and is a useful factor for determining the hardness of a problem for a GA
    - Problems exhibiting very low epistasis are most efficiently processed using a greedy algorithm
    - If a problem contains very high epistasis, then there is too little structure in the solution space, and GA will most likely drift and settle on a local optimum
    - In between the two extremes lies a type of problems suitable for GAs

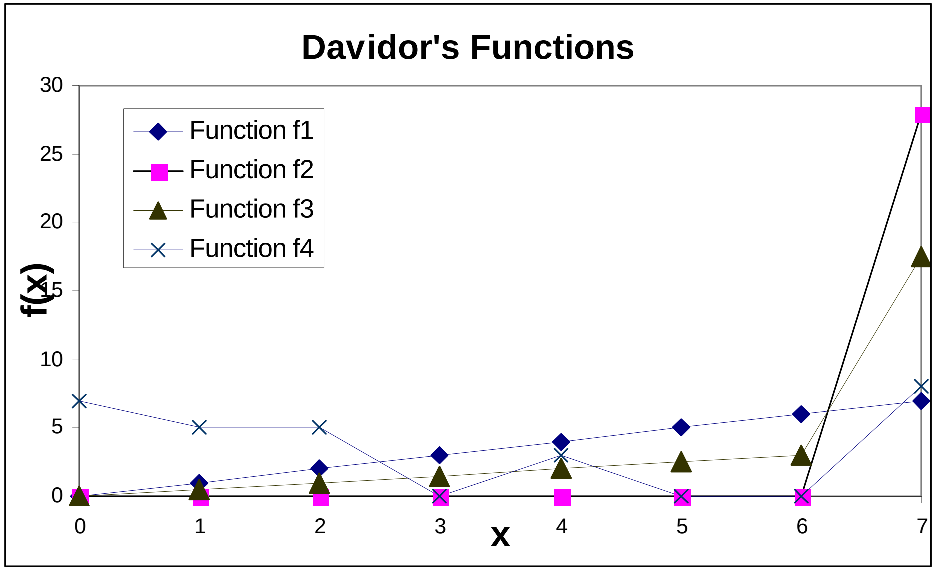
Epistasis in GAs

Design of Experiments:

Problem domains:

Simple functions defined on binary strings of length 3:

- Linear function f1
- Delta function f2
- Semi-linear function f3
- Minimal deceptive function f4 (Goldberg, 1987)



Epistasis in GAs

- Davidor's analysis indicates that:
  - Epistatic variance measure behaves as expected for linear problems
  - Increases (as it should) with qualitatively more epistatic problems
  - **But**…
    Gives hard to interpret results when only a subset of the universe is used for analysis (negative 'variance')

    THERE IS A PROBLEM!!!

    UNSOUND METHODOLOGY?

Epistasis in GAs

- Reeves & Wright used experimental design (ED) approach to analyze the same problem:
  - Full epistatic model

$$
\begin{aligned}
v(S) \quad = \quad & \text{constant} + \sum_{i=1}^{l} (\text{effect of allele at gene } i) \\
& + \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} (\text{interaction between alleles at gene } i \text{ and gene } j) \\
& + \ldots \\
& + (\text{interaction between alleles at gene } 1, \text{ gene } 2, \ldots, \text{ gene } l) \\
& + \text{random error}
\end{aligned}
$$

Epistasis in GAs

- Davidor implicitly assumed an underlying linear model (defined on bits) for the fitness of strings
  - The general model for a string with 3 binary bits:

$$v_{pqrs} = \mu + \alpha_p + \beta_q + (\alpha\beta)_{pq} + \gamma_r + (\alpha\gamma)_{pr} + (\beta\gamma)_{qr} + (\alpha\beta\gamma)_{pqr} + \varepsilon_{pqrs}$$

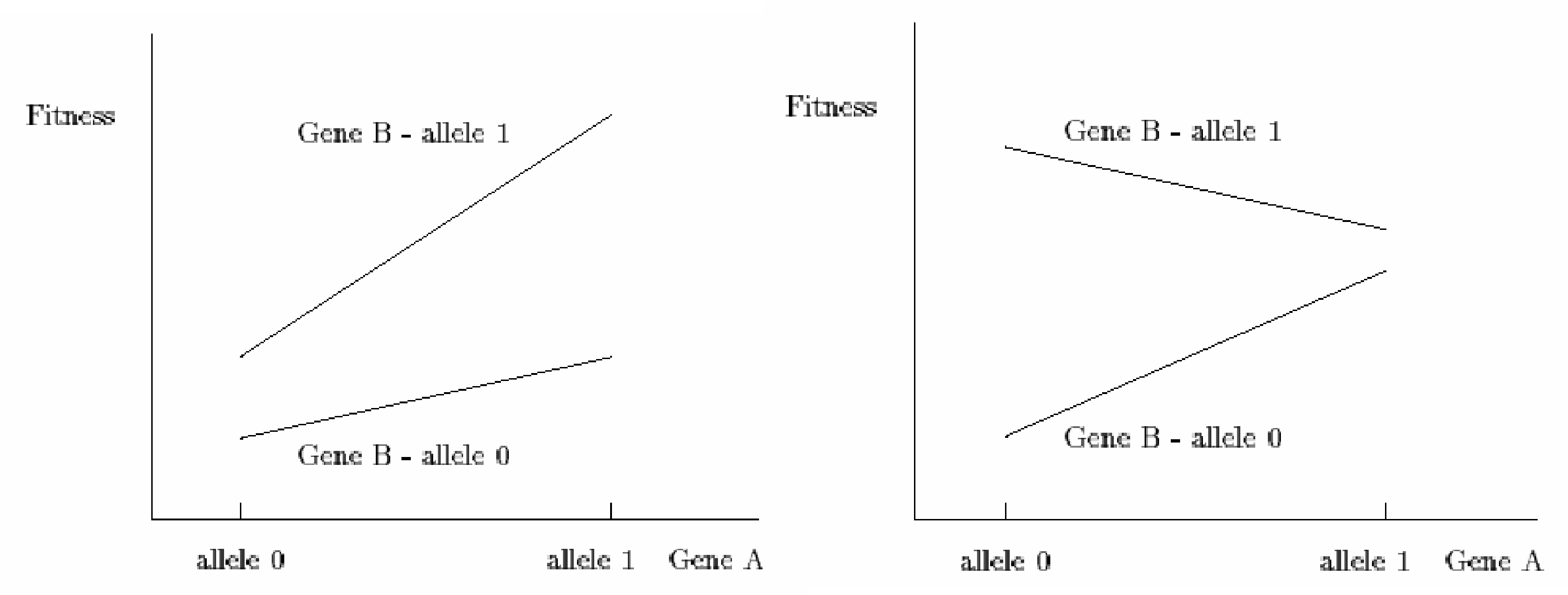
  - Davidor's model in Reeves & Wright notation corresponds to:

$$\mu + \alpha_p + \beta_q + \gamma_r$$

Hence, the epistasis measure $\epsilon(S)$ introduced by Davidor is only the sum of first-order interaction terms (higher order interactions don't contribute at all)

Epistasis in GAs

- There are various types of epistasis and not all of them contribute to the hardness of a problem for GAs:



- Did Davidor ask specific enough questions?

Epistasis in GAs

- Specific questions >> Better methodology
  >> Correct model
  >> Clearer answers
  - Explained problems with measuring epistasis given only a sample of the universe
  - Found connections of their model to Walsh functions
  - Analyzed the influence of coding on the epistasis (and directly relate their results to those obtained by Liepins and Vose)
  - Designed their own algorithm based on Sequential Elimination of Levels (SEL) method

Outline of Discussion

Part I:	Methodology	$\surd$
Part II:	Designing Experiments	$\surd$
Example:	Exp. Methodology & Design	$\surd$
Part III:	Conducting Experiments	$\longleftarrow$
Part IV:	Presenting Experiments	
Example:	Conducting & Presenting Exp.	

EC Toolkits

- Reasons to use one:
  - Save development time
  - Consistent with existing research implementations
  - Duplication is far more feasible
  - Efficient way to communicate details of implementation
- Reasons to "roll your own"
  - (More) Certain of all choices made
  - "Learning Curve" time versus development time
  - The dreaded "Work Around"
  - Mowing your lawn with a tractor
- Debugging versus experimenting
  - Validating with dual implementations
  - Duplication versus replication

Other Tools

- Statistics & Visualization
  - Be comfortable with the tool
  - Choose something others use
  - Be confident in its validity
  - Consider workflow efficiency
  - Consider production quality of graphics
- Random number generators
  - Some generators have inherent biases
  - Generators differ in sensitivity to initial seed
  - Generators differ in terms of performance
  - Generators differ in terms of length of sequence
  - EC results *can* be affected by these effects!

Tips & Tricks

- Organizing experimental groups
  - Have a top-level category for “study”, named appropriately *(e.g., "Mutation rate study")*
  - Name experimental groups with level values *(e.g., "Mutation rate experiment, Pm=0.1")*
  - Match your file & directory names to this nomenclature
- Turning around results quickly
  - Multiple passes, increasing resolution of parameter values
  - Multiple passes, increasing number of trials per group
  - Parallelism:
    - Need most results from few groups first → layer trials across machines
    - Need some results from most groups first → layer groups

Outline of Discussion

Part I:	Methodology	$\surd$
Part II:	Designing Experiments	$\surd$
Example:	Exp. Methodology & Design	$\surd$
Part III:	Conducting Experiments	$\surd$
Part IV:	Presenting Experiments	$\longleftarrow$
Example:	Conducting & Presenting Exp.	

Find the Story

- A singular driving point
  - Try to focus on one question only
  - Try to formulate the question in a clear, succinct way
  - The "story" may be different than experimental history
- A clear point
  - Don't need to include every experiment
  - Present only what is germane to the point
  - Avoid presenting experiments that confuse the point
  - *Do not omit* experiments that *weaken* the conclusion
- A replicable point
  - Provide enough detail to replicate the experiment
  - Do not overwhelm reader with tedious details
  - Can also provide accessible secondary sources

Presenting results

- Visualizing results
  - Good visualization practices *are important*
  - Have reason & purpose for presence of graphs / tables used
  - Have reason & purpose for *type* of graphs / tables used
  - Convey only relevant information! (avoid ”eye candy”)
  - Visualizations used during research *aren't necessarily the same* as those used in publication
- Presenting statistics
  - Do not claim anything empirically that you cannot defend statistically!
  - Use the correct statistical test
  - State which tests you used in a publication
  - Be careful about the word ”significant”

Suggestions and Opinions

- Suggestions
  - Distinguish clearly between what you claim to believe and what you claim to demonstrate empirically
  - If it is hard to posit a single question that captures the point of the story, it may suggest that the research questions are too vague
  - If the results do not make sense, it may suggest a problem in methodology or experimental design
- Opinions
  - If you are unconvinced, so is the audience
  - If you are convinced, the audience may still not be
  - *That* something is demonstrated empirically is nearly always less interesting than *why* it is the case:
    - Empirical presentations should have an explanatory element to them

Outline of Discussion

Part I:	Methodology	$\surd$
Part II:	Designing Experiments	$\surd$
Example:	Exp. Methodology & Design	$\surd$
Part III:	Conducting Experiments	$\surd$
Part IV:	Presenting Experiments	$\surd$
Example:	Conducting & Presenting Exp.	$\longleftarrow$

Epistasis in GAs

- Experiments comparing SEL-based algorithm with standard GA approach:
  - Problem domain:
    - Engineering design problem of a hydraulic system
    - System has 6 basic components
    - Each component has 5 types
    - Search space 5^6 = 15,625 points
  - Selecting a group of elite solutions (85) that had fitness within 15% of the overall optimum
  - Proof-of-concept problem

Epistasis in GAs

- Experimental parameters:

SEL:	GA:
Latin Square design: - initial stage: 25 points - next stage: 32 points - third stage: 27 points - last stage: 32 points -> total 116 evaluations	- The same 25 initial points form an initial population for the GA - Steady-state GA - GA run for further 91 evaluations (total of 116)

Epistasis in GAs

- Experimental parameters:

SEL:	GA:
3 flavors of the method used: - SEL-mean - SEL-max - SEL-mod (with elitism)	Representation: - string of 6 genes - each gene with 5 values Operators: Mutation rate 0.05 Unbiased uniform crossover Linear ranking selection

Epistasis in GAs

- Frequency of identification of at least one of the elite solutions (out of 100 trials):

Group	SEL *-mean*	SEL *-max*	SEL *-mod*	GA
I	1	12	93	27
II	27	25	7	25
Total	28	37	100	52

Epistasis in GAs

- Frequency of identification of at least one of the elite solutions (out of 100 trials) for non-orthogonal initial populations:

Balanced random initial population				
Group	SEL *-mean*	SEL *-max*	SEL *-mod*	GA
I	0	17	25	24
II	20	21	27	33
Total	20	38	52	57

Unbalanced random initial population				
Group	SEL *-mean*	SEL *-max*	SEL *-mod*	GA
I	4	8	9	20
II	9	18	29	26
Total	13	26	38	46

Epistasis in GAs

- The most important effects in the problem:

Effect	% variation
Main effects	
D	59%
B	6%
C	2%
F	2%
2-factor interactions	
BD	10%
DF	3%
3-factor interactions	
ADF	5%
4-factor interactions	
ABDF	3%
Total	90%

Epistasis in GAs

- Conclusions based on experimental results:
  - In general SEL approach was inferior to GA, even when orthogonal designs were used
  - One of SEL methods (SEL-mod) performed extremely well when the orthogonal designs were supplemented by elitism
  - However, even SEL-mod proved to be substantially less robust to departures from orthogonality

Epistasis in GAs

- Some interpretations of the results:
  - The approach that worked least well was SEL-mean, which works like an explicit schema-processing method
  - -> GAs seem to be doing something more than mere schema processing

References

Booth, W. C., Williams, J. M., & Colomb, G. G. (2003). The craft of research (Chicago guides to writing, editing, and publishing). Chicago, IL: University of Chicago Press.

Davidor, Y. (1990). Epistasis variance: suitability of a representation to genetic algorithms. *Complex Systems, 4*, 369-383.

Davidor, Y. (1991). Epistasis variance: a viewpoint on GA-hardness. In G. J. E. Rawlins (Ed.), *Foundations of Genetic Algorithms I* (pp. 23-35). San Mateo, CA: Morgan Kaufmann.

Goldberg, D. E. (1987). Simple genetic algorithms and the minimal deceptive problem. In L. Davis (Ed.), Genetic algorithms and simulated annealing (pp. 74-88). London: Pitman.

Leedy, P. D., & Ormrod, J. E. (2000). *Practical research: planning and design*: Prentice Hall.

Reeves, C. R., & Wright, C. C. (1995). *An experimental design perspective on genetic algorithms.* Paper presented at the Foundations of Genetic Algorithms 3, San Mateo, CA.

References

Reeves, C. R., & Wright, C. C. (1995). *Epistasis in genetic algorithms: an experimental design perspective.* Paper presented at the 6th International Conference on Genetic Algorithms (ICGA-95), Pittsburgh, PA, USA.

Reeves, C. R., & Wright, C. C. (1995). *Genetic algorithms and statistical methods: a comparison.* Paper presented at the 1st IEE/IEEE International Conference on Genetic Algorithms for Engineering Systems: Innovations and Applications, Sheffield, UK.

Reeves, C. R., & Wright, C. C. (1997). *Genetic algorithms versus experimental methods: a case study.* Paper presented at the 7th International Conference on Genetic Algorithms, East Lansing, MI, USA.

Reeves, C. R., & Wright, C. C. (1999). Genetic algorithms and the design of experiments. In D. D. Lawrence & M. D. Vose & K. A. De Jong & L. D. Whitley (Eds.), *Evolutionary Algorithms*: Springer Verlag.