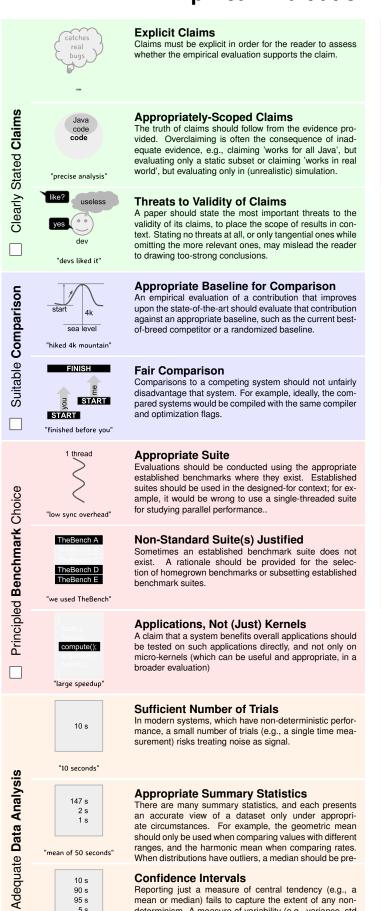
Empirical Evaluation Checklist (alpha version)

Relevant Metrics

Appropriate and Clear Experimental Design



Confidence Intervals

understand the distribution of the data.

Reporting just a measure of central tendency (e.g., a

mean or median) fails to capture the extent of any non-

determinism. A measure of variability (e.g., variance, std deviation, quantiles) and/or confidence intervals help to

10 s

90 s

5 s

"50 seconds"

\$\$\$

"energy consumed"

Direct or Appropriate Proxy Metric

If the most relevant evaluation metric is not (or cannot be) measured directly, the proxy metric used instead must be well justified. For example, a reduction in cache misses is not an appropriate proxy for actual end-to-end performance or energy consumption.

Measures All Important Effects

The costs and benefits of a technique may be multifaceted. All facets should be considered, both costs and benefits. For example, compiler optimizations may speed up programs but at the cost of drastically increasing com-



"devs were satisfied"

Sufficient Information to Repeat

Experiments should be described in sufficient detail to be repeatable. All parameters (including default values) should be included, as well as all version numbers of software, and full details of hardware platforms.

"sped up apache"

Reasonable Platform

The evaluation should be on a platform that can reasonably be said to match the claims. For example, a claim that relates to performance on mobile platforms should not have an evaluation performed exclusively on server.



"for sensor net'

Explores Key Design Parameters

Key parameters should be explored over a range to evaluate sensitivity to their settings. Examples include the size of the heap when evaluating garbage collection and the size of caches when evaluating a locality optimization.



Open Loop in Workload Generator

Load generators for transaction-oriented systems should not be gated by the rate at which the system responds. Rather, the load generator should be 'open loop', generating work independent of the performance of the system



test

19+3 22√ 7+1 8√

"perfect"

Cross-Validation Where Needed

When a system aims to be general but was developed by training on or close consideration of specific examples, it is essential that the evaluation explicitly perform crossvalidation, so that the system is evaluated on data distinct from the training set.



Comprehensive Summary Results

Experiments should be described in sufficient detail to be repeatable. All parameters (including default values) should be included, as well as all version numbers of software, and full details of hardware platforms

"have up to 4 leaves"



A truncated graph (with an axis not including zero) can exaggerate the importance of a difference. While 'zooming' in to the interesting range of an axis can potentially aid exposition, there is a significant risk that this is misleading (especially if it is not immediately clear that the axis is truncated).



Appropriate Presentation of Results

sped up B lots'

Ratios Plotted Correctly When ratios such as speedups and slowdowns are plot-

ted, the size of the bars must be linearly/logarithmically proportional to the change. When shown on the same linear scale, results are visually distorted by 1/r, where r is the ratio. This misleading effect can be avoided either by using a log scale or by normalizing to the lowest (highest)



Appropriate Level of Precision

The number of significant digits should reflect the precision of the experiment. Reporting improvements of '49.9%' when the experimental error is +/- 1% is an example of mis-stated precision, misleading the reviewer's understanding of the significance of the rest.



"9.36 s startup time'