




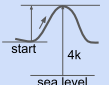






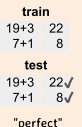










Empirical Evaluation Checklist (beta)

Clearly Stated Claims	 <p>Explicit Claims</p> <p>Claims must be explicit in order for the reader to assess whether the empirical evaluation supports the claim.</p>	Relevant Metrics	 <p>Direct or Appropriate Proxy Metric</p> <p>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.</p>
	 <p>Appropriately-Scoped Claims</p> <p>The truth of claims should follow from the evidence provided. Overclaiming is often the consequence of inadequate evidence, e.g., claiming 'works for all Java', but evaluating only a static subset or claiming 'works on real hardware', but evaluating only in (unrealistic) simulation.</p>		 <p>Measures All Important Effects</p> <p>The costs and benefits of a technique may be multi-faceted. All facets should be considered, both costs and benefits. For example, compiler optimizations may speed up programs but at the cost of drastically increasing compile times.</p>
	 <p>Threats to Validity of Claims</p> <p>A paper should state the most important threats to the validity of its claims, to place the scope of results in context. Stating no threats at all, or only tangential ones while omitting the more relevant ones, may mislead the reader to drawing too-strong conclusions.</p>		
Suitable Comparison	 <p>Appropriate Baseline for Comparison</p> <p>An empirical evaluation of a contribution that improves upon the state-of-the-art should evaluate that contribution against an appropriate baseline, such as the current best-of-breed competitor or a randomized baseline.</p>	Appropriate and Clear Experimental Design	 <p>Reasonable Platform</p> <p>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.</p>
	 <p>Fair Comparison</p> <p>Comparisons to a competing system should not unfairly disadvantage that system. For example, ideally, the compared systems would be compiled with the same compiler and optimization flags.</p>		 <p>Explores Key Design Parameters</p> <p>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.</p>
Principled Benchmark Choice	 <p>Appropriate Suite</p> <p>Evaluations should be conducted using the appropriate established benchmarks where they exist. Established suites should be used in the designed-for context; for example, it would be wrong to use a single-threaded suite for studying parallel performance.</p>	Appropriate Presentation of Results	 <p>Open Loop in Workload Generator</p> <p>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 under test.</p>
	 <p>Non-Standard Suite(s) Justified</p> <p>Sometimes an established benchmark suite does not exist. A rationale should be provided for the selection of home-grown benchmarks or subsetting established benchmark suites.</p>		 <p>Cross-Validation Where Needed</p> <p>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 cross-validation, so that the system is evaluated on data distinct from the training set.</p>
	 <p>Applications, Not (Just) Kernels</p> <p>A claim that a system benefits overall applications should be tested on such applications directly, and not only on micro-kernels (which can be useful and appropriate, in a broader evaluation)</p>		
Adequate Data Analysis	 <p>Sufficient Number of Trials</p> <p>In modern systems, which have non-deterministic performance, a small number of trials (e.g., a single time measurement) risks treating noise as signal.</p>	Appropriate Presentation of Results	 <p>Comprehensive Summary Results</p> <p>Appropriate statistics should be used to characterize the full range of results, not just the most favorable values, which may be outliers. For example, it is not appropriate to summarize speedups of 4%, 6%, 7%, and 49% as 'up to 49%'.</p>
	 <p>Appropriate Summary Statistics</p> <p>There are many summary statistics, and each presents an accurate view of a dataset only under appropriate circumstances. For example, the geometric mean should only be used when comparing values with different ranges, and the harmonic mean when comparing rates. When distributions have outliers, a median should be presented.</p>		 <p>Axes Include Zero</p> <p>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).</p>
	 <p>Report Data Distribution</p> <p>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 understand the distribution of the data.</p>		 <p>Ratios Plotted Correctly</p> <p>When ratios such as speedups and slowdowns are plotted, 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) value.</p>
			 <p>Appropriate Level of Precision</p> <p>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.</p>