Multi-Seed Benchmark Consistency