

Table 1: Computations on the \mathbb{X} dataset: p -values for FILO2^x vs FILO2 on the left and FILO2^x (long) vs FILO2 (long) on the right.

p -values in bold are associated with rejected hypothesis when $\bar{\alpha} = 0.002778$.

The last column of each group contains a p -value interpretation. In particular, FILO2^x is not statistically different from FILO2 when H_0 cannot be rejected (Similar), FILO2^x is statistically better when both H_0 and H_1 are rejected (Better), and, finally, FILO2^x is statistically worse when H_0 is rejected and H_1 is not rejected (Worse).

Solvers	FILO2 ^x vs FILO2		FILO2 ^x (long) vs FILO2 (long)	
	H_0 / H_1	Outcome	H_0 / H_1	Outcome
2	0.636031 / 0.318015	Similar	0.686502 / 0.343251	Similar
3	0.861598 / 0.430799	Similar	0.986721 / 0.493360	Similar
4	0.204705 / 0.897648	Similar	0.918803 / 0.540598	Similar
5	0.400299 / 0.799850	Similar	0.748675 / 0.374337	Similar
6	0.093292 / 0.953354	Similar	0.702061 / 0.351031	Similar
7	0.019983 / 0.990009	Similar	0.352392 / 0.823804	Similar
8	0.354791 / 0.822604	Similar	0.184298 / 0.907851	Similar
9	0.146985 / 0.926507	Similar	0.434540 / 0.782730	Similar
10	0.022747 / 0.988626	Similar	0.112887 / 0.943557	Similar