Table 1: Computations on the $\mathbb{B}$ 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).

| | FILO2^x vs FILO2 | | FILO2^x (long) vs FILO2 (long) | |
Solvers	H_0 / H_1	Outcome	H_0 / H_1	Outcome
2	0.375000 / 0.838867	Similar	0.695312 / 0.687500	Similar
3	0.556641 / 0.753906	Similar	0.695312 / 0.687500	Similar
4	0.492188 / 0.784180	Similar	0.556641 / 0.753906	Similar
5	0.845703 / 0.615234	Similar	0.625000 / 0.312500	Similar
6	0.232422 / 0.903320	Similar	0.921875 / 0.577148	Similar
7	0.695312 / 0.687500	Similar	0.322266 / 0.161133	Similar
8	0.695312 / 0.687500	Similar	1.000000 / 0.539062	Similar
9	0.083984 / 0.967773	Similar	0.625000 / 0.312500	Similar
10	0.375000 / 0.838867	Similar	0.769531 / 0.652344	Similar