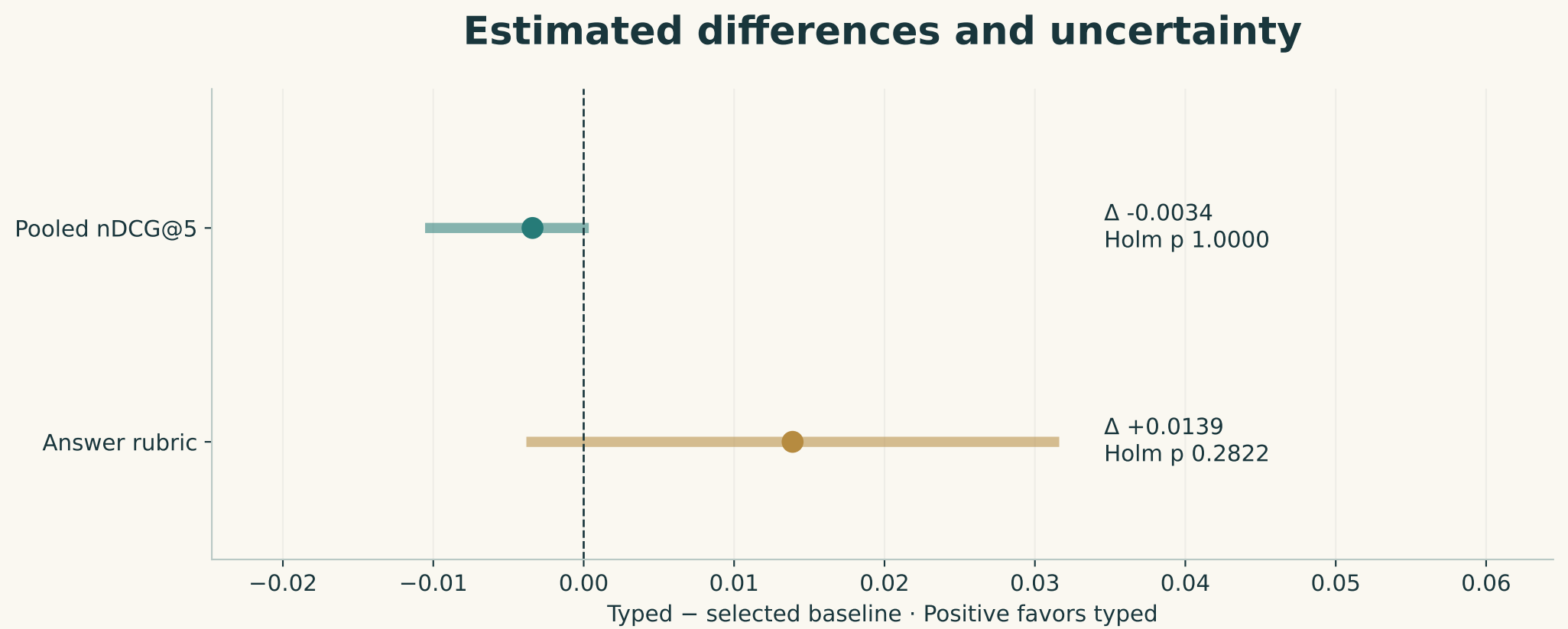

Estimated differences and uncertainty
Pooled nDCG@5
Answer rubric
Δ -0.0034
Holm p 1.0000
Δ +0.0139
Holm p 0.2822
−0.02
−0.01
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Typed − selected baseline · Positive favors typed
95% paired percentile bootstrap intervals · 10,000 resamples · Two co-primary tests