Stability of Dataset Comparisons for wig+100 (p=0.1) 0.35 **Dataset Pair** beir-webis-touche2020 vs msmarco-passage beir-webis-touche2020 vs msmarco-subsample 0.30 msmarco-passage vs msmarco-subsample 0.25 Swap Probability 0.20 0.15 0.10 0.05 0.00 3 5 Average Performance Delta