Stability of Dataset Comparisons for wig+100 (p=0.5) **Dataset Pair** 0.40 beir-webis-touche2020 vs msmarco-passage beir-webis-touche2020 vs msmarco-subsample 0.35 msmarco-passage vs msmarco-subsample 0.30 0.25 Swap Probability 0.20 0.15 0.10 0.05 0.00 Average Performance Delta