Stability of Dataset Comparisons for wig+100 (p=0.9) **Dataset Pair** • 0.175 beir-webis-touche2020 vs msmarco-passage beir-webis-touche2020 vs msmarco-subsample msmarco-passage vs msmarco-subsample 0.150 0.125 Swap Probability 0.100 0.075 0.050 0.025 0.000 Average Performance Delta