Stability of Dataset Comparisons for scq (p=0.9) **Dataset Pair** beir-webis-touche2020 vs msmarco-passage 0.06 beir-webis-touche2020 vs msmarco-subsample msmarco-passage vs msmarco-subsample 0.05 0.04 0.03 0.02 0.01 0.00 25 50 75 100 125 Average Performance Delta

Swap Probability