Stability of Dataset Comparisons for scq (p=0.5) 0.30 **Dataset Pair** beir-webis-touche2020 vs msmarco-passage beir-webis-touche2020 vs msmarco-subsample msmarco-passage vs msmarco-subsample 0.25 0.20 0.15 0.10 0.05 0.00 25 50 75 100 125 Average Performance Delta

Swap Probability