Accuracy under different training data size and different test difficulty