Test loss for different model sizes