95% credible interval is well calibrated across replications