Supplemental Materials

Supplement 1: Iterations and convergence

In this supplement, we identified the minimum iterations needed in the Bayesian-GRM, so that the posterior samples can provide a stable estimate of each parameter. For each parameter in the Bayesian-GRM, including the discrimination parameters α, the population intercept parameters C, the ED severity parameters θ, and the logistic regression parameters, we examined whether posterior iterations can provide stable 5% and 95% quantiles. When we run multiple independent Bayesian samplings, and the estimates from each of these samplings have a less than 0.05 standard deviation, we consider these samplings to provide a stable estimate of parameters.

After testing different numbers of iterations using our empirical data, we found that a minimum of 4,000 iterations could provide stable estimates of all parameters. In the SCOFF, we ran ten chains independently, each containing 500 warm-up samples and 4,000 iterations. All parameters have standard deviations less than 0.05 in their 5% quantiles and 95% quantiles across ten samplings, except for 16 person parameters θ measuring individual-level ED severity. The standard deviations corresponding to these parameters were less than 0.1.

Similarly, we ran ten chains independently for the BASE, each containing 500 warm-up samples and 4,000 iterations. All parameters have standard deviations less than 0.05 in their estimated 5% and 95% quantiles.

For both the SCOFF and the BASE, we validated that these chains had reasonable convergence based on the $\hat{R}$ statistic (Gelman & Rubin, 1992), which were less than 1.05 for all parameters. The effective sample size per chain exceeded 1,000 for all parameters.

As a result, we consider 4,000 iterations to be a reasonable size for moderately-sized datasets using either the SCOFF or the BASE. We inflated this number to 6,000 iterations in this study.

Supplement 2: Comparing models with invariant and variant discrimination parameters

In the main study, we used a Bayesian-GRM that had invariant discrimination parameters α across groups. The assumption underlying this model is that items in the ED screening surveys have an invariant ability to discriminate participants, regardless of whether these participants are cisgender men, cisgender women, or non-cisgender individuals. We compare this model with a variant version of the GRM where different discrimination parameters α are used for men, women, and non-cisgender individuals. We used the leave-one-out (LOO) cross-validation method (Vehtari et al., 2017) to evaluate these models. The leave-one-out information criterion (LOOIC) is used to compare these models, where a model with a lower LOOIC is a superior model. We computed LOOIC using the R package "loo" (Vehtari et al., 2021).

When applied to the SCOFF data, LOOIC for the invariant model was -2502.9, and the LOOIC for the variant model was -2499.2. This indicates that the variant model is slightly better for the SCOFF data than the invariant model. When applied to the BASE data, LOOIC for the invariant model was -2275.2, and the LOOIC for the variant model was -2295.3. This indicates that the invariant model is better for the BASE data than the variant model.

These model comparisons provide mixed results on whether the variant or the invariant model would be superior to fit ED screening data. Considering that ED screening studies usually have a small to moderate sample size, making it more challenging to estimate parameters in the variant model than the invariant model, we used the invariant model for this study.

Supplement 3: Parameter recovery using simulation

We used a simulation study to examine the parameter recovery ability of the Bayesian-GRM. Specifically, we would like the model to recover the data generating parameters accurately when the simulated data

had a relatively small sample size.

We performed 50 independent simulations for each ED screener (the SCOFF and the BASE). In each simulation, our simulated data included 90 cisgender women, 60 cisgender men, and 30 individuals identified as non-cisgender, resulting in a total of 180 participants. To generate each simulated dataset, we selected the following set of true data-generating parameters: the ED severity parameters θ were randomly drawn from all estimated posterior parameters from the corresponding screener, when the model was fitted to the observed data. The other parameters were sampled from the posterior distributions generated by fitting the Bayesian-GRM to the observed data. Using each set of parameters, we simulated survey responses and individual ED diagnosis.

After obtaining 50 simulated datasets for each ED screener, we fit the Bayesian GRM to each dataset using "rstan", and evaluated whether the posterior parameters can recover the true parameters used to generate simulated data.

We then examined whether true parameter values were contained in the 95% equal-tailed credible intervals of the estimated parameters from simulated data. For the SCOFF, the true data generating discriminability parameters (α) were within the 95% equal-tailed credible intervals in 90%-100% of simulations. The true person-parameters θ were within the 95% equal-tailed credible intervals in 95.6% cases. The true population intercept parameters C were within 95% equal-tailed CIs in 92%-100% of simulations. The intercept parameters of the logistic regression, b_o, b_f, and b_m, were within 95% CIs in 96%, 90%, and 90% of simulations, respectively. Lastly, the slope parameters of the logistic regression, $b_{o,1}$, $b_{f,1}$, and $b_{m,1}$, were within 95% CIs in 76%, 80%, and 94% of simulations, respectively.

For the BASE, the true data generating discriminability parameters (α) were within the 95% equal-tailed credible intervals in 88%-100% of simulations. The true person-parameters θ were within the 95% equal-tailed credible intervals in 95.7% cases. The true population intercept parameters C were within 95% equal-tailed CIs in 90%-100% of simulations. The intercept parameters of the logistic regression, b_o, b_f, and b_m, were within 95% CIs in 98%, 90%, and 70% of simulations, respectively. Lastly, the slope parameters of the logistic regression, $b_{o,1}$, $b_{f,1}$, and $b_{m,1}$, were within 95% CIs in 90%, 98%, and 70% of simulations, respectively.

Parameter recovery results for the GRM parameters were overall satisfactory. However, for the groups of cisgender men and individuals identified as non-cisgender, the parameter recovery results for the logistic parameters were sub-optimal. Considering that the parameter recovery results were better for the group of cisgender women (N=90), we used the sample size of 90 for both cisgender women and cisgender men in the bootstrap study in the main manuscript.

Supplement 4: Small sample performance - Results from every bootstrapped sample

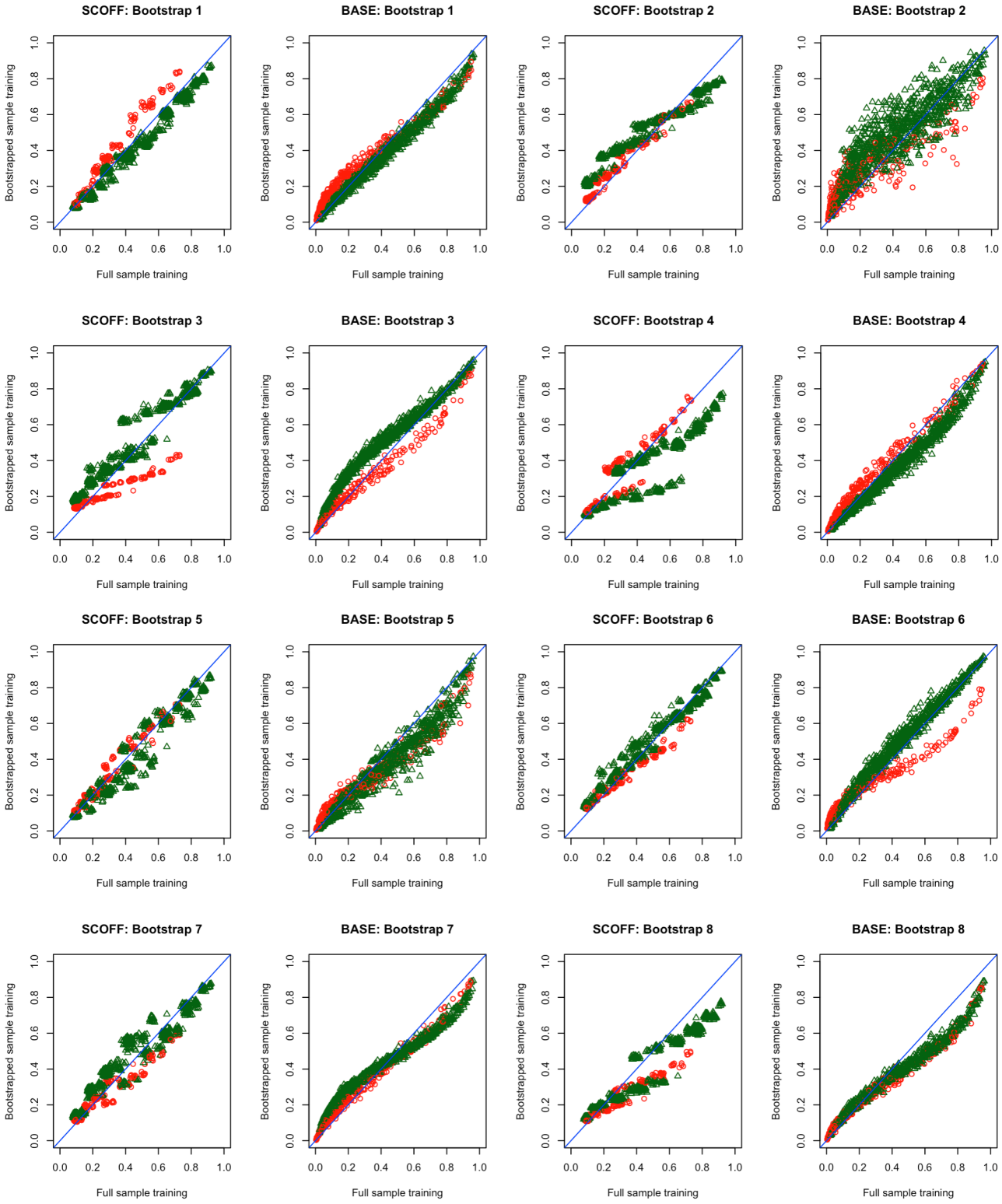
In the main manuscript, we showed the bootstrap result from the first bootstrapped sample in *Figure 8*. We display results from the other bootstrapped samples below in this Supplement.

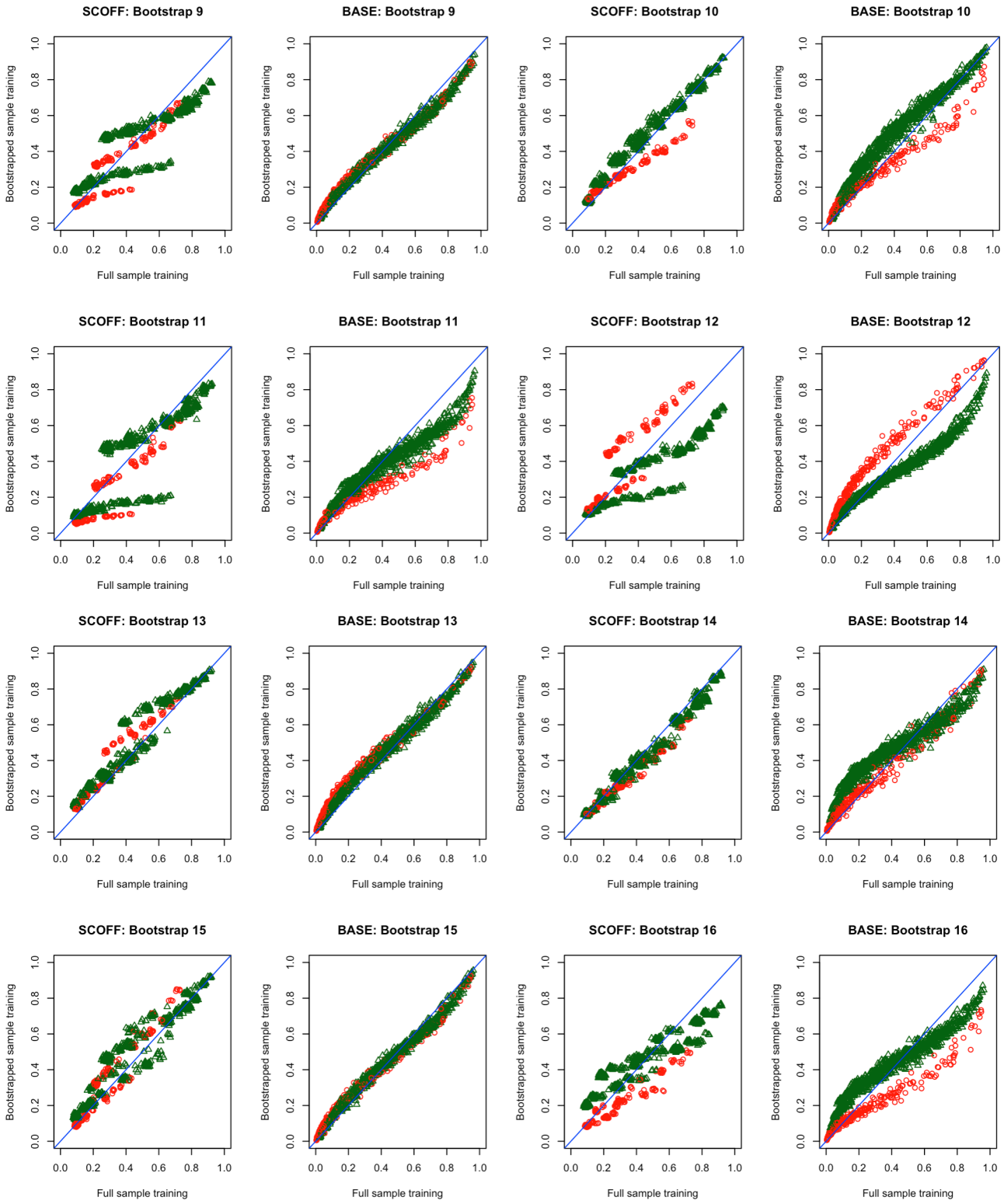
References

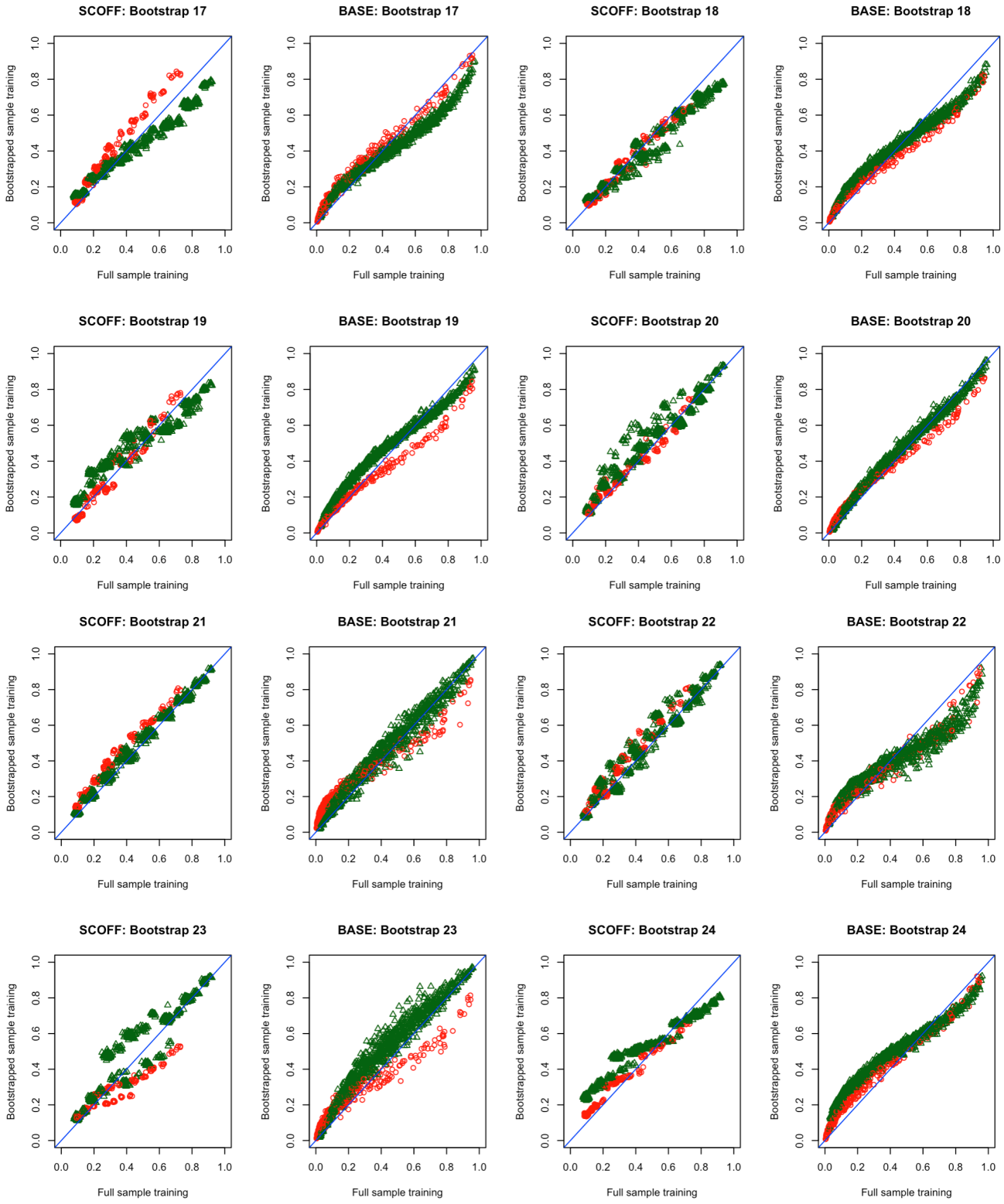
Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 457–472.

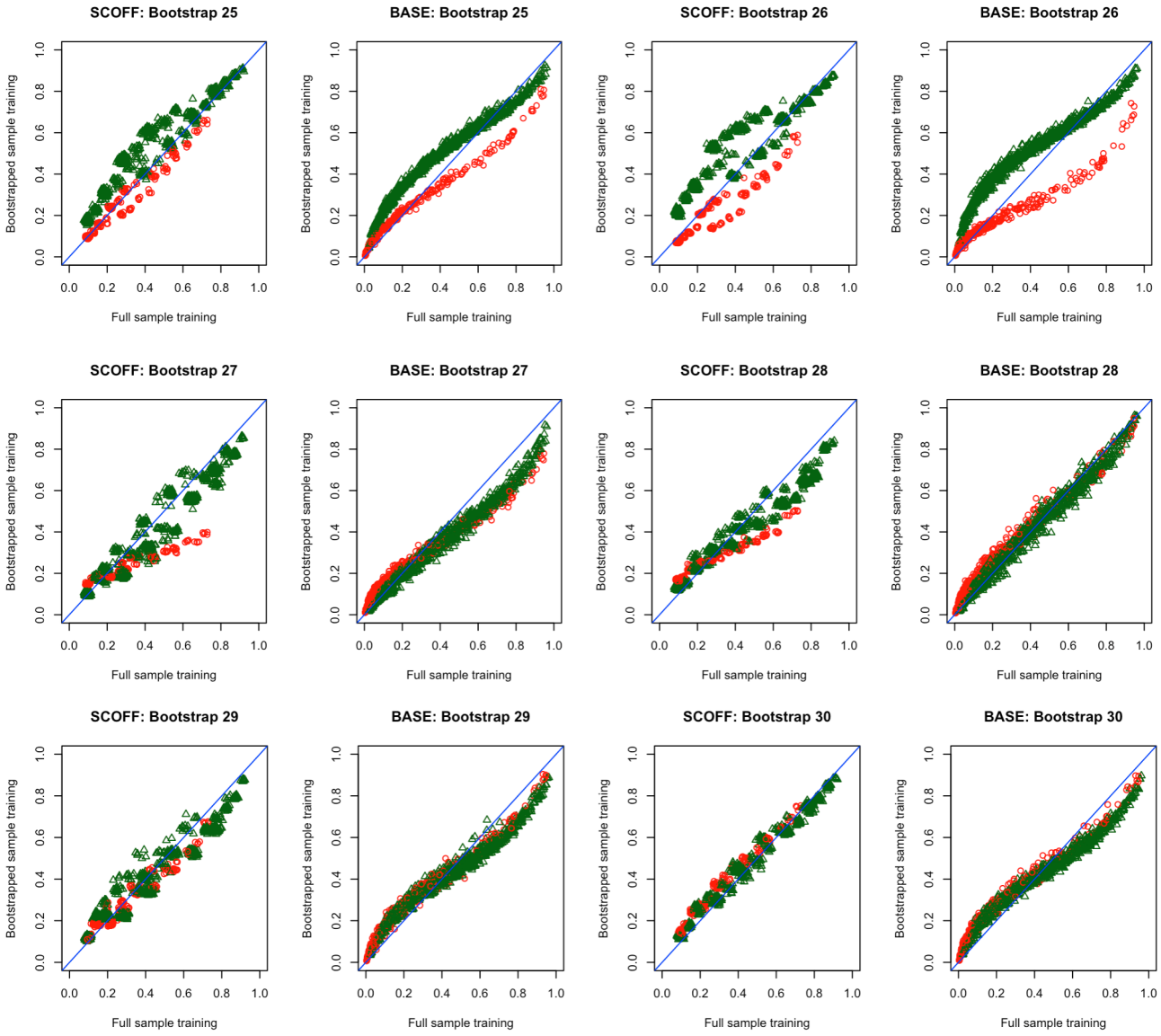
Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, *27*, 1413–1432.

Vehtari, A., Gelman, A., Gabry, J., & Yao, Y. (2021). Package 'loo'. *Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models*.


SCOFF: Bootstrap 1
BASE: Bootstrap 1
SCOFF: Bootstrap 2
BASE: Bootstrap 2
SCOFF: Bootstrap 3
BASE: Bootstrap 3
SCOFF: Bootstrap 4
BASE: Bootstrap 4
SCOFF: Bootstrap 5
BASE: Bootstrap 5
SCOFF: Bootstrap 6
BASE: Bootstrap 6
SCOFF: Bootstrap 7
BASE: Bootstrap 7
SCOFF: Bootstrap 8
BASE: Bootstrap 8
Bootstrapped sample training
Full sample training
0.0
0.2
0.4
0.6
0.8
1.0


SCOFF: Bootstrap 9
BASE: Bootstrap 9
SCOFF: Bootstrap 10
BASE: Bootstrap 10
SCOFF: Bootstrap 11
BASE: Bootstrap 11
SCOFF: Bootstrap 12
BASE: Bootstrap 12
SCOFF: Bootstrap 13
BASE: Bootstrap 13
SCOFF: Bootstrap 14
BASE: Bootstrap 14
SCOFF: Bootstrap 15
BASE: Bootstrap 15
SCOFF: Bootstrap 16
BASE: Bootstrap 16
Bootstrapped sample training
Full sample training
0.0
0.2
0.4
0.6
0.8
1.0


SCOFF: Bootstrap 17
BASE: Bootstrap 17
SCOFF: Bootstrap 18
BASE: Bootstrap 18
SCOFF: Bootstrap 19
BASE: Bootstrap 19
SCOFF: Bootstrap 20
BASE: Bootstrap 20
SCOFF: Bootstrap 21
BASE: Bootstrap 21
SCOFF: Bootstrap 22
BASE: Bootstrap 22
SCOFF: Bootstrap 23
BASE: Bootstrap 23
SCOFF: Bootstrap 24
BASE: Bootstrap 24
Bootstrapped sample training
Full sample training
0.0
0.2
0.4
0.6
0.8
1.0


SCOFF: Bootstrap 25
BASE: Bootstrap 25
SCOFF: Bootstrap 26
BASE: Bootstrap 26
SCOFF: Bootstrap 27
BASE: Bootstrap 27
SCOFF: Bootstrap 28
BASE: Bootstrap 28
SCOFF: Bootstrap 29
BASE: Bootstrap 29
SCOFF: Bootstrap 30
BASE: Bootstrap 30
Bootstrapped sample training
Full sample training
0.0
0.2
0.4
0.6
0.8
1.0