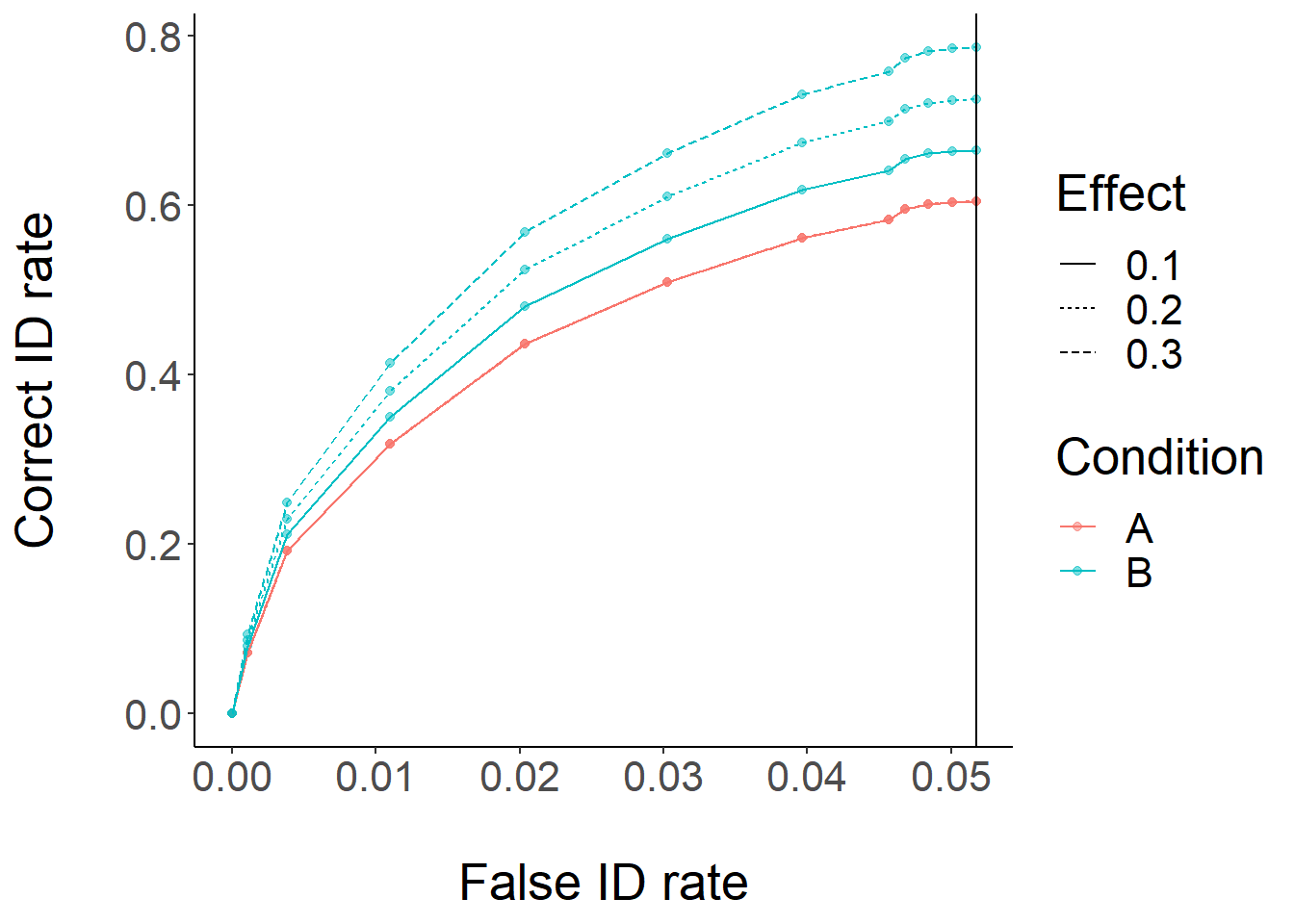
**Supplementary Material**

**Supplementary Material A: Example summary report**

## **Simulation parameters**

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| Ns | 500, 1000, 1500 |
| Effects | .1, .2, .3 |
| Between- or within-subjects | Between-subjects |
| Confidence levels | 11 |
| Effects by confidence level | N/A |
| # of lineups/subject | 2 |
| # TA lineups/subject | 1 |
| # TP lineups/subject | 1 |
| # of simulated samples per effect size/N | 100 |
| # of bootstraps per AUC/DPP test | 3000 |
| Partial AUC truncation | Lowest false ID rate |
| Custom specificity | NA |
| Two-tailed or one-tailed? | 2\_tail |
| Type I error rate | 0.05 |

## **Plot of hypothetical ROC curves that were tested**

****

## 

## 

## 

## 

## 

## 

## 

## 

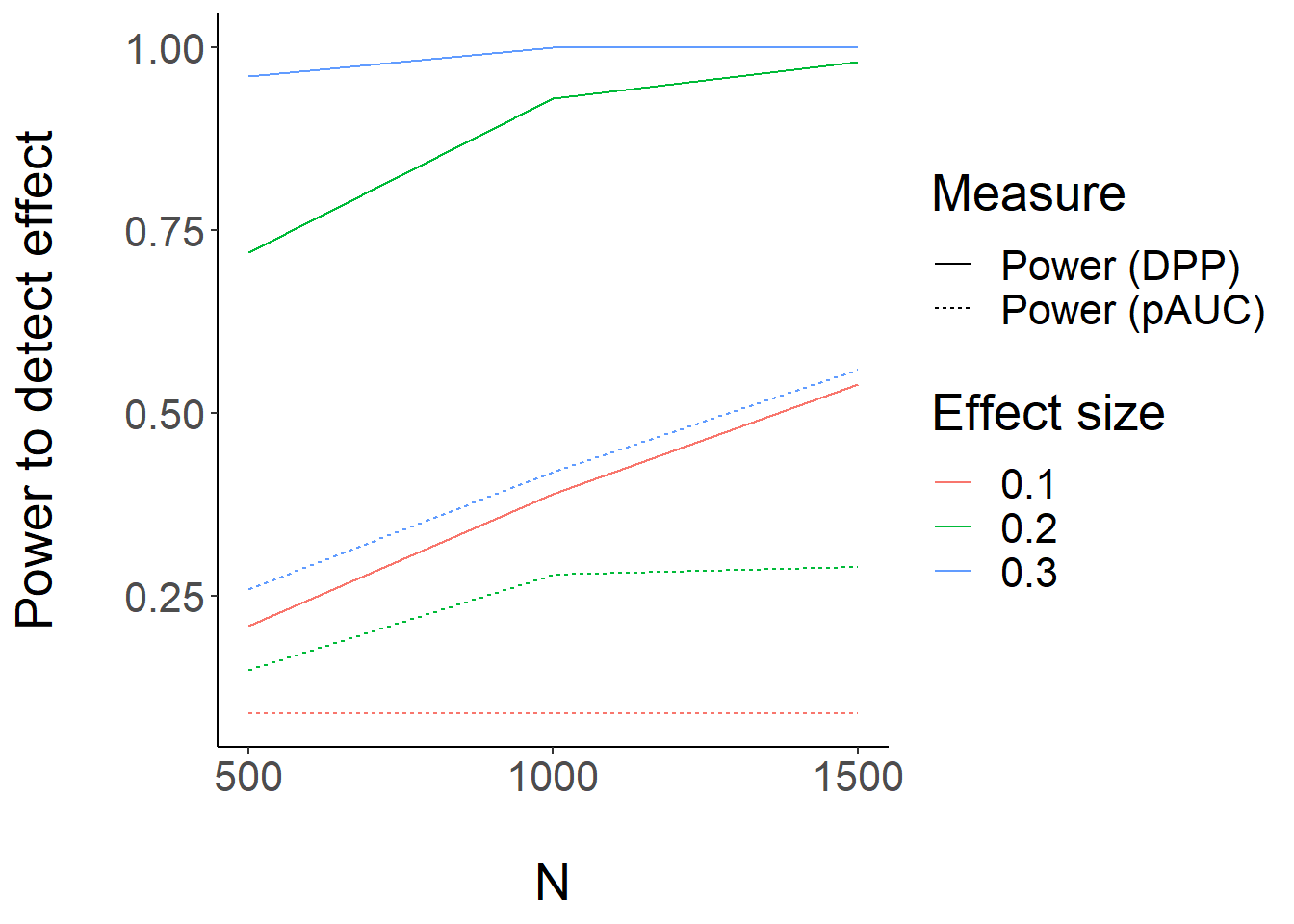
## 

## **Simulation results**

### **Time taken: 1056.51 minutes**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **N** | **Effect size** | **Power (pAUC)** | **Power (DPP)** | **Avg. AUC in A** | **AUC A 95% CI lower** | **AUC A 95% CI upper** | **Avg. AUC in B** | **AUC B 95% CI lower** | **AUC B 95% CI upper** | **Avg. AUC difference** | **AUC difference 95% CI lower** | **AUC difference 95% CI upper** | **Avg. DPP in A** | **DPP A 95% CI lower** | **DPP A 95% CI upper** | **Avg. DPP in B** | **DPP B 95% CI lower** | **DPP B 95% CI upper** | **Avg. DPP difference** | **DPP difference 95% CI lower** | **DPP difference 95% CI upper** |
| 500 | 0.1 | 0.09 | 0.21 | 0.017851 | 0.006215 | 0.029454 | 0.019968 | 0.008055 | 0.032776 | -0.00212 | -0.00996 | 0.005306 | 0.566029 | 0.511027 | 0.6075 | 0.524113 | 0.472655 | 0.575836 | 0.041916 | -0.02294 | 0.114191 |
| 1000 | 0.1 | 0.09 | 0.39 | 0.019709 | 0.012653 | 0.027841 | 0.02141 | 0.013102 | 0.029853 | -0.0017 | -0.00727 | 0.004202 | 0.5702 | 0.541746 | 0.594132 | 0.525453 | 0.48405 | 0.563164 | 0.044747 | 0.000705 | 0.088773 |
| 1500 | 0.1 | 0.09 | 0.54 | 0.020277 | 0.015804 | 0.027438 | 0.021856 | 0.013538 | 0.029681 | -0.00158 | -0.00666 | 0.003375 | 0.568262 | 0.536052 | 0.59913 | 0.523844 | 0.493839 | 0.555358 | 0.044418 | 0.006418 | 0.088227 |
| 500 | 0.2 | 0.15 | 0.72 | 0.017664 | 0.008748 | 0.026505 | 0.020884 | 0.010618 | 0.032301 | -0.00322 | -0.01296 | 0.005718 | 0.569415 | 0.516473 | 0.612582 | 0.47988 | 0.437209 | 0.525164 | 0.089535 | 0.011582 | 0.148609 |
| 1000 | 0.2 | 0.28 | 0.93 | 0.019556 | 0.010982 | 0.026956 | 0.023616 | 0.014863 | 0.032303 | -0.00406 | -0.0094 | 0.003009 | 0.571775 | 0.532268 | 0.605705 | 0.479809 | 0.446846 | 0.509568 | 0.091966 | 0.034496 | 0.134909 |
| 1500 | 0.2 | 0.29 | 0.98 | 0.020126 | 0.014093 | 0.026992 | 0.023809 | 0.014191 | 0.032281 | -0.00368 | -0.00831 | 0.001437 | 0.570256 | 0.539897 | 0.603221 | 0.478747 | 0.449473 | 0.506667 | 0.091509 | 0.048118 | 0.130618 |
| 500 | 0.3 | 0.26 | 0.96 | 0.017943 | 0.005323 | 0.030745 | 0.02432 | 0.009497 | 0.03938 | -0.00638 | -0.01573 | 0.002486 | 0.571716 | 0.518382 | 0.638473 | 0.427422 | 0.381418 | 0.477618 | 0.144295 | 0.066164 | 0.2273 |
| 1000 | 0.3 | 0.42 | 1 | 0.019054 | 0.011657 | 0.026043 | 0.024657 | 0.014468 | 0.034621 | -0.0056 | -0.01096 | 0.000738 | 0.567326 | 0.532536 | 0.606382 | 0.430542 | 0.400032 | 0.455818 | 0.136784 | 0.089627 | 0.185768 |
| 1500 | 0.3 | 0.58 | 1 | 0.019869 | 0.013956 | 0.026117 | 0.026037 | 0.016859 | 0.034546 | -0.00617 | -0.01199 | -0.00046 | 0.570504 | 0.537327 | 0.599206 | 0.43033 | 0.40237 | 0.454409 | 0.140175 | 0.098003 | 0.176867 |

### **Power curves**

****

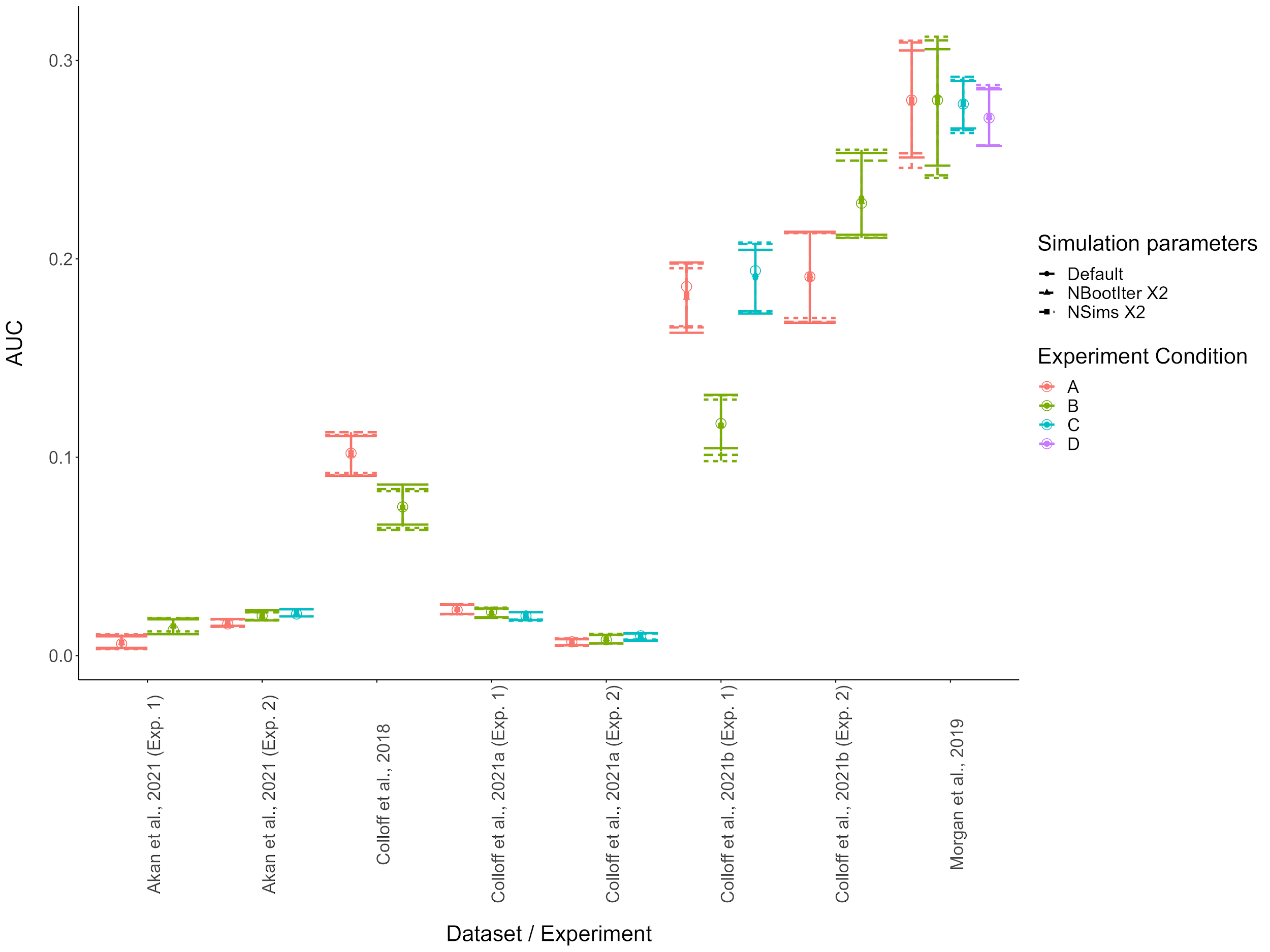
**Supplementary Material B: App validation & testing results**

# **powe(R)OC Testing Results**

This document details several tests of simulation validity/performance conducted using real datasets. Data/articles used for testing can all be found in the “References” section below and at<https://github.com/E-Y-M/poweROC/tree/main/Dataset%20testing%20and%20reports> , and were obtained from the Open Science Framework. If you have ROC data (along with analysis parameters) you are willing to share for the purposes of simulation testing, feel free to email me at ericmah@uvic.ca. Issues/comments on the app or simulation testing results can be posted on GitHub at<https://github.com/E-Y-M/poweROC/issues>.

## **AUC Recovery**

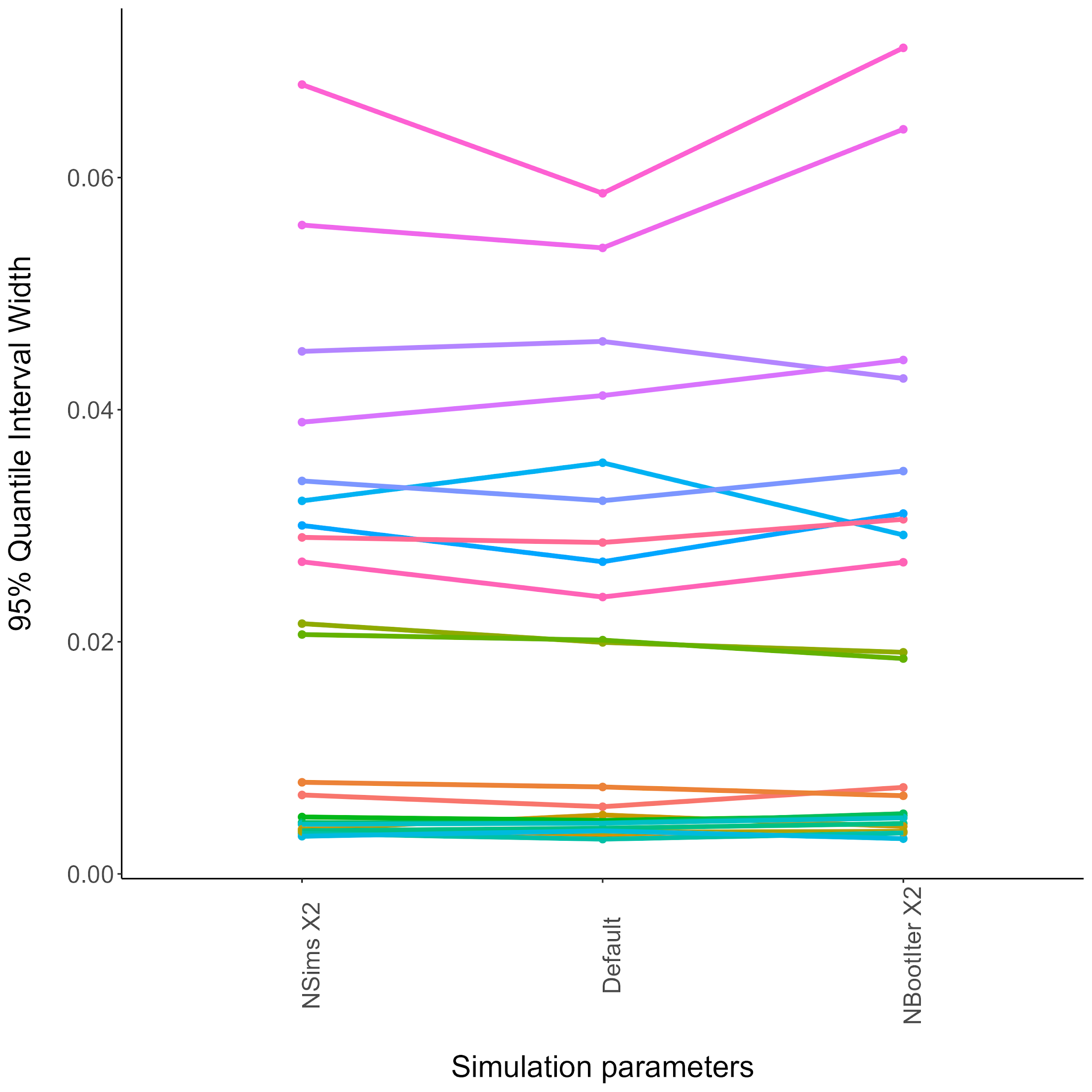
At a basic level, simulation validity depends on the ability of the simulations to recover AUC values close to those in the original dataset. The figure below depicts original AUC estimates from various papers with open data (5 papers, 8 experiments, 22 ROC curves computed using the same N’s/pAUC cutoffs in the original papers), along with simulated estimates and intervals:



Testing the ability of the simulation to recover AUC values from experiments. Open circles represent original AUC values, all other points represent simulation estimates under various conditions (“NSims” = Number of simulated datasets per effect size/N, “NBootIter” = Number of bootstrap iterations per AUC comparison). Error bars = 95% quantiles on the mean estimated AUC for the simulations. Overall, simulations demonstrate excellent ability to recover original AUC values, even under default settings (NSims = 100, NBootIter = 1000).

## **Simulation precision under different conditions**

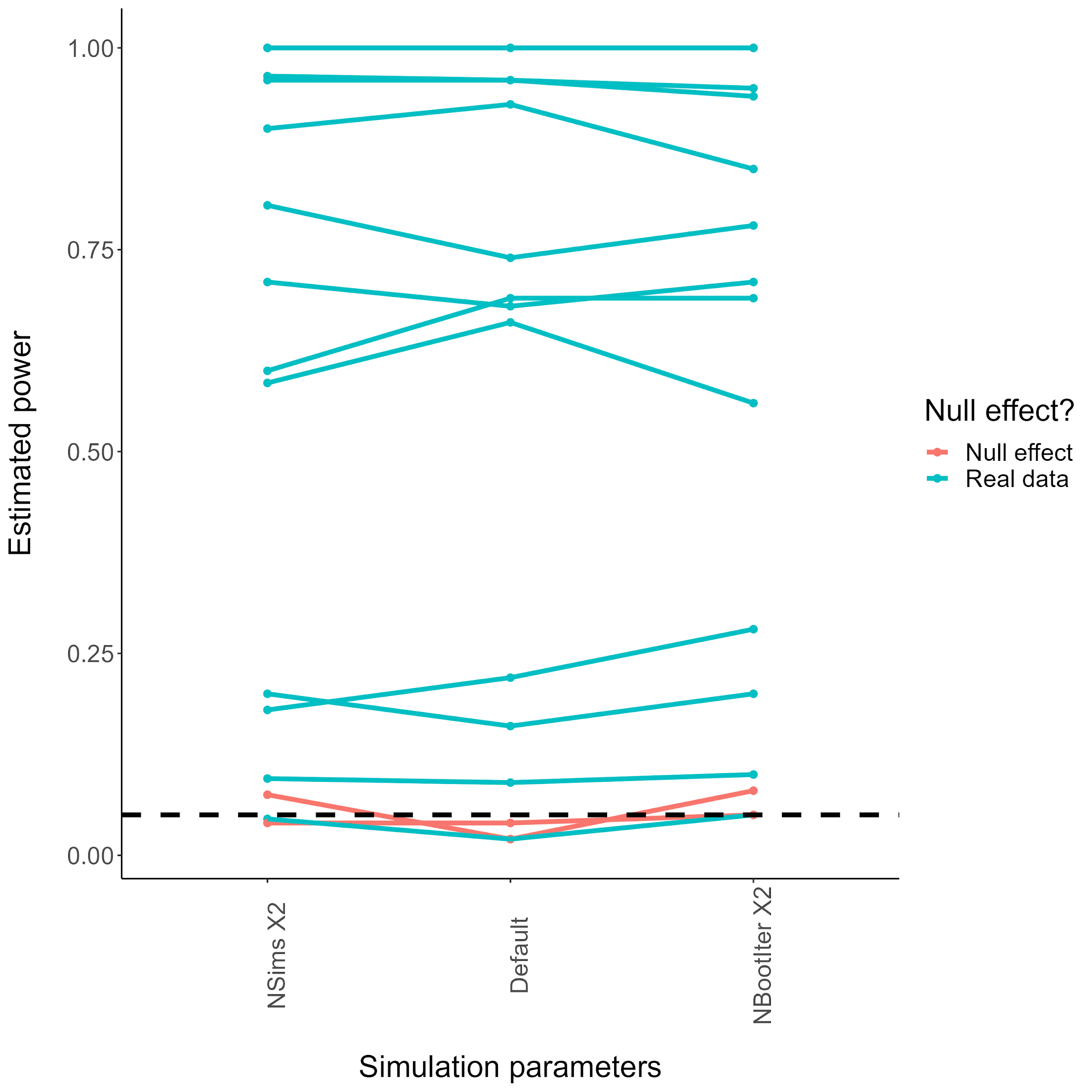
Still, the question remains as to whether increasing the number of simulations or bootstrap iterations increases power. The figure below shows the width of the 95% quantile intervals for the AUC estimates above, as a function of the simulation conditions.



Based on these simulations, it does not seem that increasing the number of simulations or bootstrap iterations necessarily or substantially increases the precision of the simulation beyond the default settings, suggesting that the default settings will result in reasonable estimates.

## **Power estimates under different conditions**

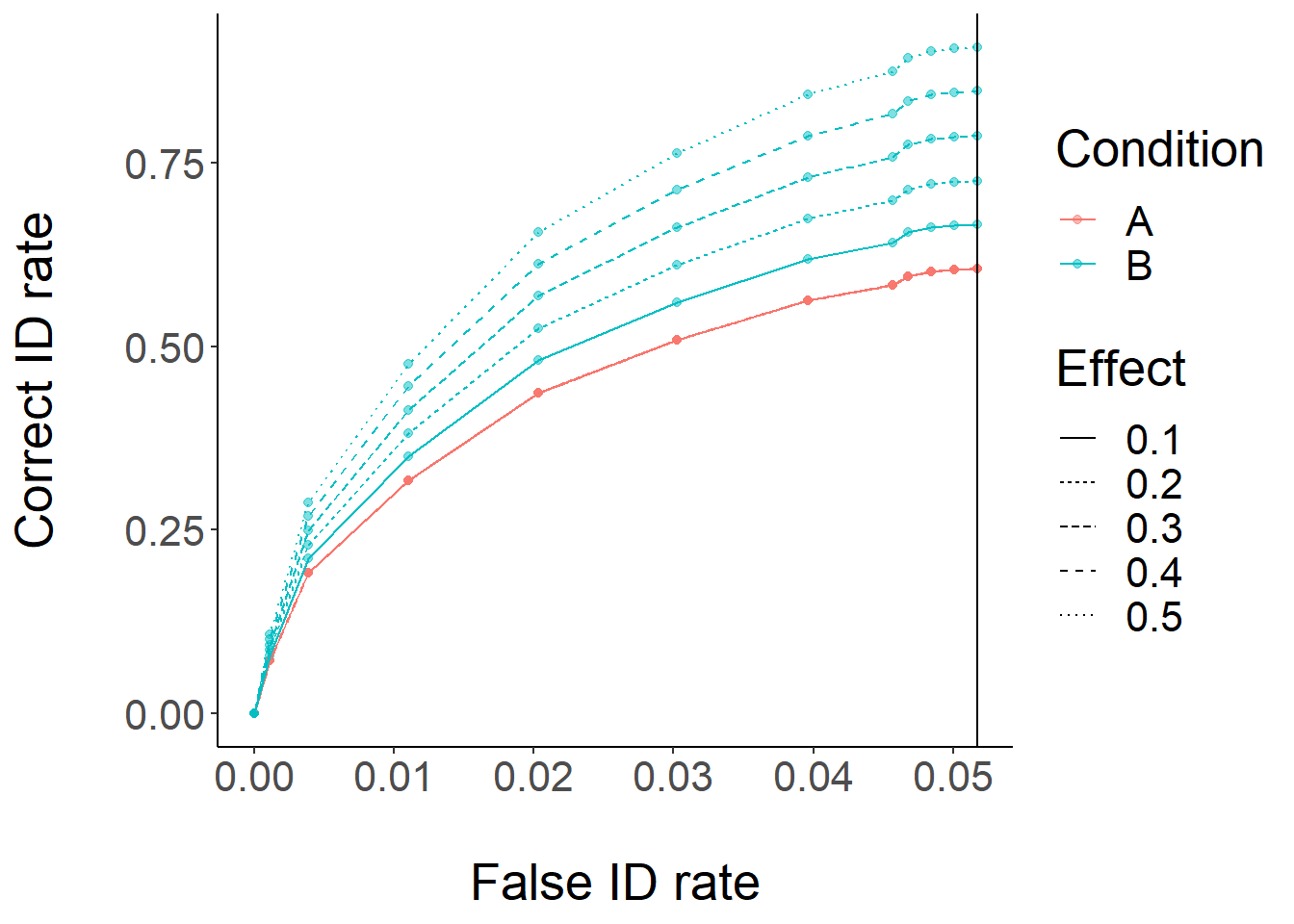
It is not clear whether the default simulation settings result in the most accurate power estimates. I simulated power for 13 ROC comparisons from the papers above. I also conducted two simulation runs of a dataset with a prespecified null effect (using the “Medium Similarity” condition from Colloff et al., 2021a as a base) to compare with the normative Type I Error Rate of .05, all under the three different simulation conditions. These power estimates are plotted below:



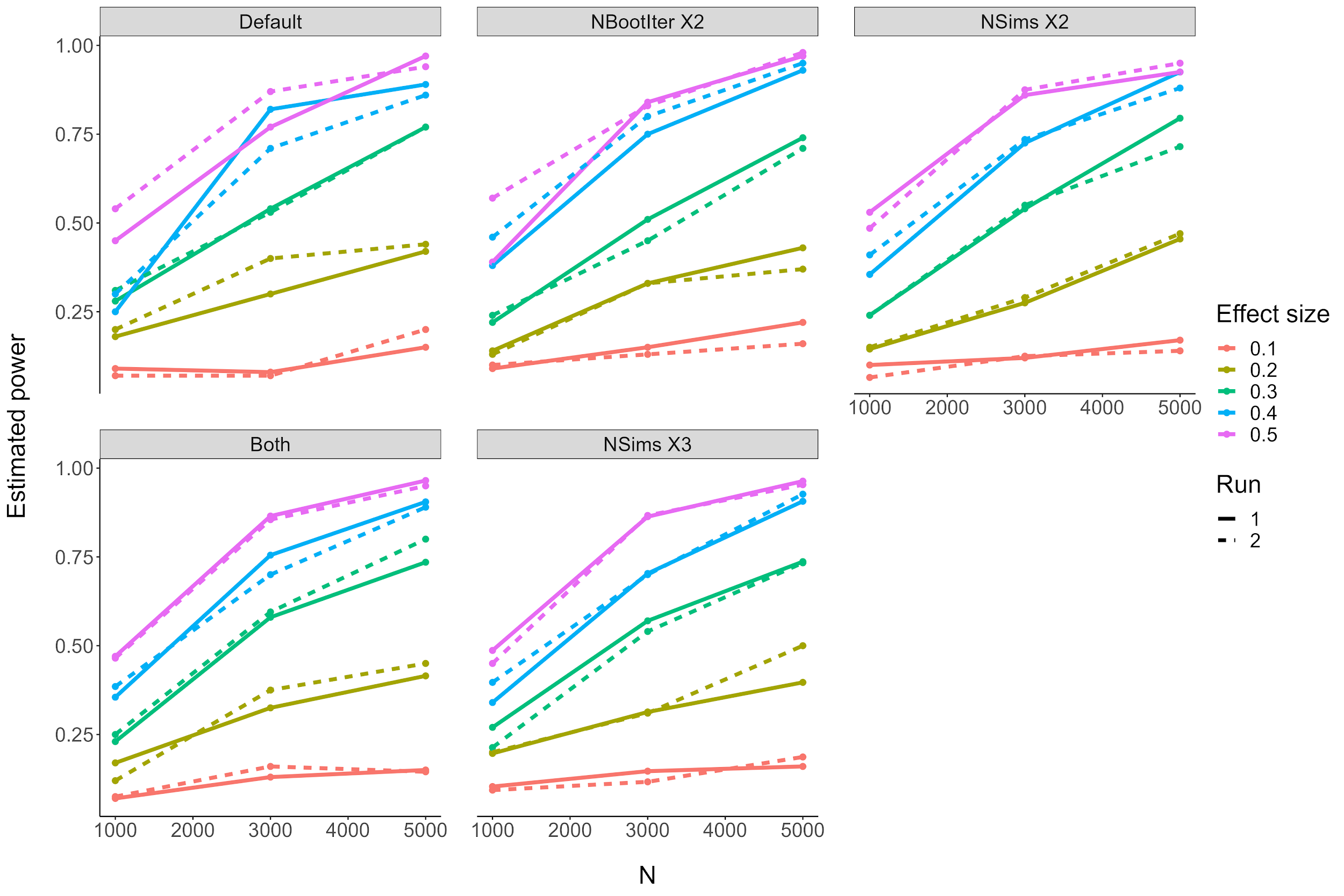
Power estimates differed slightly across the different simulation conditions, but no clear patterns emerged. In these examples, the maximum range of estimated power was .10. Importantly, power estimates in the two null effect simulations were close to the nominal Type I Error Rate of .05 (though the non-default settings resulted in slightly higher estimates).

## **Power curves in a full simulation example**

Finally, I examined the behaviour of the different simulation settings for a full simulation example (i.e., involving multiple N’s/sample sizes). I simulated power for 5 effect sizes and 3 sample sizes (1000, 3000, 5000), again using the “Medium Similarity” condition data from Colloff et al. (2021a) as a base. For each simulation setting I ran two simulations to get a basic idea of run-to-run consistency. First, the hypothetical ROCs that were tested for this analysis:



Next, power curves for these simulations:



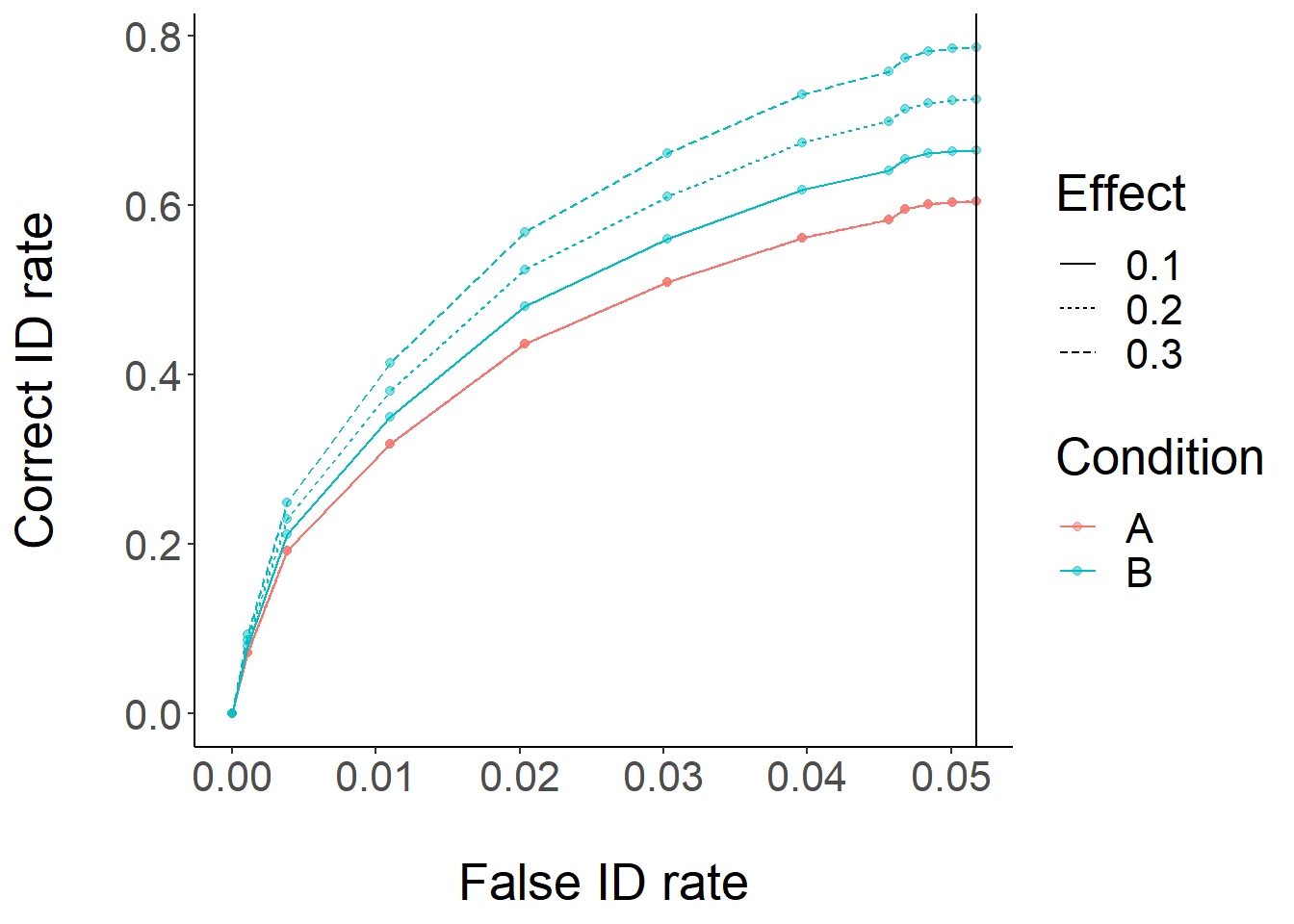
These simulations result in the same general expected pattern, but there are a few things worth noting. First, it appears that the default settings (though the fastest for simulation) result in a good amount of run-to-run variability, and a violation of power simulation expectations (i.e., higher power for a smaller effect size with the same sample size). Between increasing the bootstrap iterations and increasing the # of sims, increasing the # of sims seems to result in more run-to-run consistency while maintaining the expected patterns of results, at the cost of increasing the required simulation time. At least in these examples, upping the default values of both NSims and NIter did not seem to offer substantial benefit over increasing NSims, and increasing NSims beyond 200 did not seem to result in a substantial gain.

## **DPP testing**

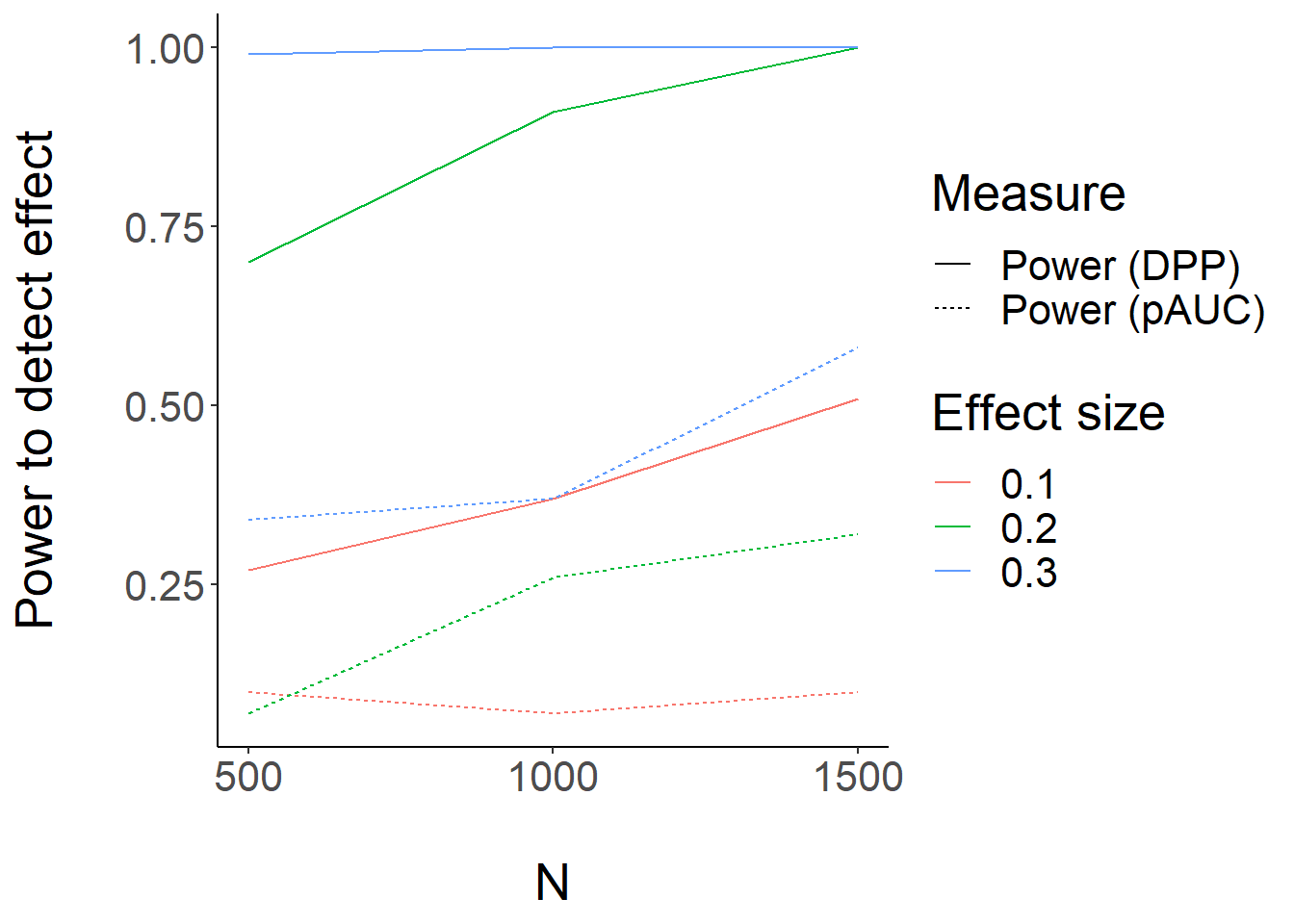
I tested a) the ability of the app to recover published DPP values (from Smith et al., 2019, who used the “concealment” and “nothing” condition data from Colloff et al., 2016). The reported DPP values in Smith et al., 2019 were .86 and .82 respectively, with a DPP difference of .04 (95% CI [.007, .087]). In an initial simulation run using the same data and sample size (and using 100 simulated samples and 2200 bootstraps per DPP test), the simulation DPPs were .87 and .82, with an estimated DPP difference of .05 (95% CI [-.008, .10]). In a second simulation run, the simulation DPPs were .87 and .82, with an estimated DPP difference of .05 (95% CI [-.006, .11]). Aside from some discrepancies with the DPP difference CIs due to slightly different calculation methods, powe(R)OC recovered the DPP values accurately and consistently across different simulation runs.

In a test of long-run Type I error rates, across two simulations using base data with a null effect the DPP Type I error rate was .04 and .03 (slightly outperforming pAUC, which had a Type I error rate of .07 in both cases).

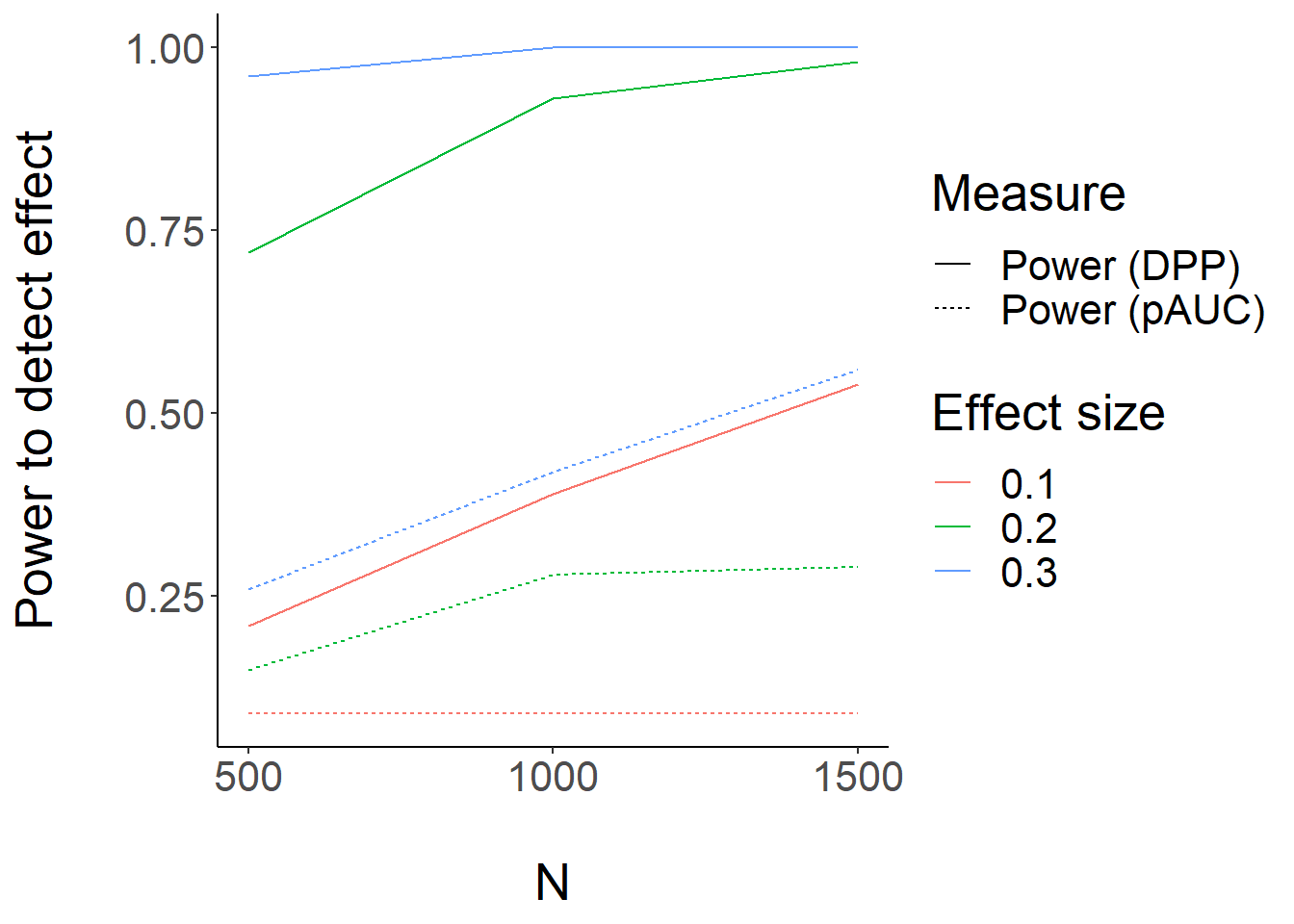
Finally, I conducted two runs of a full power simulation with both AUC and DPP. The hypothetical ROC curves were again constructed using the “Medium Similarity” condition data from Colloff et al. (2021a) as a base:



The resulting power curves for Run 1:



…and for Run 2:



Again, power estimates differed slightly across runs (mostly at smaller sample sizes), but results were consistent overall. Interestingly, in this case it appears that power to detect the differences is substantially higher for DPP than for pAUC. I hesitate to draw any general conclusions, given that this is a single simulation (and additionally, in the validation using the data from Colloff et al., 2016, pAUC held a power advantage over DPP). However, this does suggest that power can differ substantially depending on the measure used, and suggests some further avenues for testing when one measure provides more power than the other.

## **Recommendations for users**

In light of these testing results, I recommend that users: a) Use the default simulation parameters if analysis time is a concern, but to up the # of simulations per sample/effect size to 200 if time is not a concern, b) select only a few effect sizes/sample sizes for simulation (e.g., based on prior results documented in the app), c) set final planned sample size slightly higher than their target power (e.g., + .05-.10), and d) conduct a couple simulation runs (e.g., one with default settings to get a general idea of the sample size, then a finer-grained simulation including only a few sample sizes and using more simulations & bootstrap iterations).

## **References**

–Akan, M., Robinson, M. M., Mickes, L., Wixted, J. T., & Benjamin, A. S. (2021). The effect of lineup size on eyewitness identification. Journal of Experimental Psychology: Applied, 27(2), 369–392.<https://doi.org/10.1037/xap0000340>

–Carlson, C. A., & Carlson, M. A. (2014). An evaluation of lineup presentation, weapon presence, and a distinctive feature using ROC analysis. Journal of Applied Research in Memory and Cognition, 3(2), 45–53.<https://doi.org/10.1016/j.jarmac.2014.03.004>

–Colloff, M. F., Seale-Carlisle, T. M., Karoğlu, N., Rockey, J. C., Smith, H. M. J., Smith, L., Maltby, J., Yaremenko, S., & Flowe, H. D. (2021). Perpetrator pose reinstatement during a lineup test increases discrimination accuracy. Scientific Reports, 11(1), 13830.<https://doi.org/10.1038/s41598-021-92509-0>

–Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair Lineups Make Witnesses More Likely to Confuse Innocent and Guilty Suspects. Psychological Science, 27(9), 1227–1239.<https://doi.org/10.1177/0956797616655789>

–Colloff, M. F., Wade, K. A., Strange, D., & Wixted, J. T. (2018). Filler-Siphoning Theory Does Not Predict the Effect of Lineup Fairness on the Ability to Discriminate Innocent From Guilty Suspects: Reply to Smith, Wells, Smalarz, and Lampinen (2018). Psychological Science, 29(9), 1552–1557.<https://doi.org/10.1177/0956797618786459>

–Colloff, M. F., Wilson, B. M., Seale-Carlisle, T. M., & Wixted, J. T. (2021). Optimizing the selection of fillers in police lineups. Proceedings of the National Academy of Sciences, 118(8), e2017292118.<https://doi.org/10.1073/pnas.2017292118>

–Dobolyi, D. G., & Dodson, C. S. (2013). Eyewitness confidence in simultaneous and sequential lineups: A criterion shift account for sequential mistaken identification overconfidence. Journal of Experimental Psychology: Applied, 19(4), 345–357.<https://doi.org/10.1037/a0034596>

–Gronlund, S. D., Carlson, C. A., Neuschatz, J. S., Goodsell, C. A., Wetmore, S. A., Wooten, A., & Graham, M. (2012). Showups versus lineups: An evaluation using ROC analysis. Journal of Applied Research in Memory and Cognition, 1(4), 221–228.<https://doi.org/10.1016/j.jarmac.2012.09.003>

–Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. Journal of Experimental Psychology: Applied, 18(4), 361–376.<https://doi.org/10.1037/a0030609>

–Morgan, D. P., Tamminen, J., Seale-Carlisle, T. M., & Mickes, L. (2019). The impact of sleep on eyewitness identifications. Royal Society Open Science, 6(12), 170501.<https://doi.org/10.1098/rsos.170501>

–Seale-Carlisle, T. M., & Mickes, L. (2016). US line-ups outperform UK line-ups. Royal Society Open Science, 3(9), 160300.<https://doi.org/10.1098/rsos.160300>

–Smith, A. M., Lampinen, J. M., Wells, G. L., Smalarz, L., & Mackovichova, S. (2019). Deviation from Perfect Performance Measures the Diagnostic Utility of Eyewitness Lineups but Partial Area Under the ROC Curve Does Not. Journal of Applied Research in Memory and Cognition, 8(1), 50–59.<https://doi.org/10.1016/j.jarmac.2018.09.003>

–Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A., & Carlson, C. A. (2015). Effect of retention interval on showup and lineup performance. Journal of Applied Research in Memory and Cognition, 4(1), 8–14.<https://doi.org/10.1016/j.jarmac.2014.07.003>