# Bayesian inference for the Probability of Superiority

Abstract

Certain methodologists have recommended the probability of superiority (PS) as an effect size for two-group comparisons, because the PS is relatively easy to understand. In this paper, we develop a Bayesian estimation approach for the PS, and provide prior specifications for model parameters. Our approach exploits both the relationship between the PS and the Mann Whitney $U$-statistic, and the connection between parametric and nonparametric statistic. Using a simulation study, we show the approach to be efficient relative to other estimation approaches for a sample size that is small but common in educational interventions. Hence, it is our hope that the method we present helps applied researchers better communicate group differences when comparing two independent samples.

*Keywords:* effect size, probability of superiority, Bayesian estimation, independent samples t-test, rankit

Word count: 1827

Bayesian inference for the Probability of Superiority

The independent-samples $t$-test is a well-established statistical procedure for the comparison of two groups. Kruschke (2013) demonstrated that utilizing Bayesian parameter estimation for this simple procedure yields a wealth of information that the frequentist approach fails to provide. Kruschke (2013) also demonstrated how to compute Cohen's $d$, a very standard effect size for two-sample scenario. The probability of superiority (PS) is an alternative effect size that is relatively easy for nonstatisticians to interpret (Brooks, Dalal, & Nolan, 2014; McGraw & Wong, 1992). In the remainder of this paper, we review commonplace estimation methods for the PS. Next, we show how the relation of the PS to the Mann Whitney $U$-statistic provides an alternative approach for estimating the PS. This allows us to develop a Bayesian model for the PS, and we provide prior choice recommendations. Finally, we use a simulation study to compare the Bayesian approach with prevailing estimation methods when the sample size is small.

## Estimation of the probability of superiority

The PS is the probability that a randomly selected score from one group exceeds a randomly selected score from another group (McGraw & Wong, 1992). McGraw and Wong (1992) called the PS the *common language effect size* because they expected that nonstatisticians will easily understand the PS, and there is some empirical evidence to support this view (Brooks et al., 2014). We recommend the McGraw and Wong (1992) and Brooks et al. (2014) papers for readers who wish to learn more about the PS from a conceptual perspective. Assuming the data are normally distributed,
$\text{PS} = \Phi(\mu_2 - \mu_1 \mid 0, \sqrt{\sigma_2^2 + \sigma_1^2})$, where $\Phi(\cdot)$ is the normal cumulative distribution function, $\mu_j$ and $\sigma_j$ for $j = \{1,\ 2\}$ represent group means and standard deviations respectively (McGraw & Wong, 1992). If the data are normal and the group mean difference is zero, the PS will be 50%, meaning that half the time, values from one group exceed the other and vice-versa.

Vargha and Delaney (2000) gave an alternative formulation that does not make distributional assumptions of the data:

$$\widehat{\mathrm{PS}} = [\#(\mathbf{y}_2 > \mathbf{y}_1) + 0.5 \times \#(\mathbf{y}_2 = \mathbf{y}_1)]/(n_1 n_2) \tag{1}$$

where $\#(\cdot)$ is the count function and $n_1$ and $n_2$ are the sample sizes of groups 1 and 2 respectively. In equation 1, we compare each data point in a group with all the data points in the other group using the ordinal information in the data. For inference, results from Ruscio and Mullen (2012) suggest that the bias-corrected and accelerated (BCa) bootstrap performs adequately under a wide variety of data conditions.

**Relation to Mann Whitney *U*-statistic and estimation.** The Mann-Whitney $U$-test is a common nonparametric analog to the $t$-test. The $U$-test tests whether values from one group are greater than values from the other group, hence, when the null hypothesis of the $U$-test is true, the PS is 50% (Conover, 1999, p. 274; Reiczigel, Zakariás, & Rózsa, 2005). The formula for $U$ is: $\#(\mathbf{y}_2 > \mathbf{y}_1) + 0.5 \times \#(\mathbf{y}_2 = \mathbf{y}_1)$, and $U$ is the numerator in the PS formula by Vargha and Delaney (2000). One can leverage this connection between both statistics to conduct inference on the PS (Ruscio & Mullen, 2012; Vargha & Delaney, 2000).

Conover and Iman (1981) showed that one can obtain nonparametric results by applying parametric methods to rank transformed data. We exploit this approach to lay out a method for estimating the PS, which begins by ranking the data. There are many procedures for ranking data (Conover & Iman, 1981). We prefer *normal scores* using the *rankit* procedure for two reasons. Penfield and McSweeney (1968) recommended normal scores to psychologists on the basis of their efficiency. Moreover, the rankit procedure normalizes the ranks, such that the empirical density of rankits is close to the standard normal likelihood (Soloman & Sawilowsky, 2009). Rankits are: $\mathbf{y}^* = \Phi^{-1}\Big((\mathbf{r} - 0.5)/N \mid 0, 1\Big)$, where $\Phi^{-1}(\cdot)$ is the normal quantile function, $\mathbf{r}$ are the ranked data, and $N$ is the total

sample size. To compute the PS, we utilize the formulation for the PS by McGraw and Wong (1992) on the rankit statistics: $\widehat{\text{PS}} = \Phi(\hat{\mu}_2^* - \hat{\mu}_1^* \mid 0, \sqrt{\hat{\sigma}_2^{*2} + \hat{\sigma}_1^{*2}})$, where $^*$ denotes rankit statistics. We expect this normal approximation to be adequate for rankits, since rankits will be approximately standard normal, unless a considerable proportion of the data are tied.

## A Bayesian model for the probability of superiority

We will assume the rankits are Gaussian with means and standard deviations that depend on group membership, $\mathbf{y}_j^* \sim \mathcal{N}(\alpha^* + \beta_j^*,\ \sigma_j^*)$. The average of the group means is $\alpha^*$, $\beta_j^*$ captures the deviation of each group $j$ from $\alpha^*$ and $|\beta_2^* - \beta_1^*|$ is the difference between the groups. Since we are using rankits, we have considerable prior information about the data. The rankits will have a mean that is close to zero, and a standard deviation that is about one. Additionally, when all the values in group 2 are greater than the values in group 1, i.e. PS = 100%, the data in each group will be a different truncated normal distribution. Accordingly, the means of groups 1 and 2 are $-\phi\Big(\Phi^{-1}(p \mid 0, 1)\Big)/p$ and $\phi\Big(\Phi^{-1}(p \mid 0, 1)\Big)/(1-p)$ respectively, where $\phi(\cdot)$ is the standard normal density function and $p$ is the proportion of cases in group 1.[1] As an example, when ties are limited in the data and $p = 50\%$, $\max(|\beta_2^* - \beta_1^*|) \approx 1.60$ and $\max(|\alpha^*|) \approx 0$. As the level of imbalance in group sizes increases, $\max(|\beta_2^* - \beta_1^*|)$ and $\max(|\alpha^*|)$ increase as shown in Figure 1.

We also computed $|\alpha^*|$ and $|\beta_2^* - \beta_1^*|$ for varying levels of $p$ when Cohen's $d = 1$. We consider $d = 1$ to be at the limit of what is plausible or believable for educational interventions related to achievement; the 99th percentile of effect sizes from such educational interventions is about Cohen's $d = 0.9$ (Table 1 in Kraft, 2018). When Cohen's $d \leq 1$, $\alpha^*$ will largely fall in the $(-0.5,\ 0.5)$ interval and $|\beta_2^* - \beta_1^*|$ will largely fall in the $[0,\ 1)$ interval, see Figure 1.

---

[1] Based on the equation for the mean of truncated normal data.

We now make prior choice recommendations. We follow a subjective Bayesian approach (Goldstein, 2006) in the manner prescribed by Greenland (2006). Greenland (2006) recommended using the middle 95% interval of prior distributions to a-priori identify the plausible values for parameters. Given the plausible range for $\alpha^*$, we expect a prior with zero mean and a standard deviation that is a quarter of most of the data, such that $\Pr(-.5 < \alpha^* < .5) = 95\%$. And $\alpha^* \sim \mathcal{N}(0,\ 0.25)$ satisfies this expectaion. We expect one $\beta_j^*$ will be positive and the other negative. And given the plausible range for $|\beta_2^* - \beta_1^*|$, we expect $\Pr(-0.5 < \beta_j^* < 0.5) = 95\%$, hence, $\beta_j^* \sim \mathcal{N}(0,\ 0.25)$ suffices. Finally, we expect that $\Pr(0 < \sigma_j^* < 1.5) = 95\%$; hence, $\sigma_j^* \sim \mathcal{N}^+(0, 0.75)$ suffices. This is a reasonable expectation given that rankits will have a standard deviation that is about 1, but one of the $\sigma_j^*$ values may be somewhat higher. Our complete model is:

$$\begin{aligned} &\mathbf{y}_j^* \sim \mathcal{N}(\alpha^* + \beta_j^*,\ \sigma_j^*) \\ &\alpha^* \sim \mathcal{N}(0,\ s_\alpha = 0.25),\ \beta_j^* \sim \mathcal{N}(0,\ s_\beta = 0.25),\ \sigma_j^* \sim \mathcal{N}^+(0,\ s_\sigma = 0.75) \end{aligned} \tag{2}$$

Researchers may modify $s_\alpha$, $s_\beta$ or $s_\sigma$ based on context-specific information. The PS is: $\Phi(\mu_2^* - \mu_1^* \mid 0, \sqrt{\sigma_2^{*2} + \sigma_1^{*2}})$. We use the posterior samples of $\mu_2^*$, $\mu_1^*$, $\sigma_2^*$ and $\sigma_1^*$ to compute the posterior samples for the PS.

## Method

We conducted a simulation study to test the estimation quality of our Bayesian approach. We compared the Bayesian approach against the formula of McGraw and Wong (1992) applied to rankits (hereafter *rankits*), the parametric approach of McGraw and Wong (1992) (hereafter *parametric*), and the non-parametric approach of Vargha and Delaney (2000), hereafter *nonparametric*. For non-Bayesian approaches, we performed the BCa bootstrap with 2000 replications for inference.

We used Stan (Carpenter et al., 2017) as our Bayesian computation engine. Stan uses the No-U-Turn sampler (NUTS). NUTS, an algorithm for sampling continuous parameters, is more efficient than the Gibbs sampler (Hoffman & Gelman, 2014). For Bayesian parameter estimation, we drew 2000 posterior samples across 4 parallel chains, and retained the final half of the samples within each chain. This left us with 4000 samples for inference. As a crude check of sampling convergence, we computed the potential scale reduction factor ($\widehat{R}$) and effective sample size ($n_{\mathsf{eff}}$) for the posterior samples of the PS. $\widehat{R}$ close to 1 and $n_{\mathsf{eff}}$ above a thousand are preferred (Carpenter et al., 2017).

We performed 500 iterations within each design condition. The total sample size within each iteration was always 70; a small sample size but one that is common enough in educational interventions (Slavin & Smith, 2009). To test estimation quality of the different methods, we produced a chart showing the distribution of the estimated PS marking out the mean, interquartile range, 5th, and 95th percentiles. We also marked out the true population parameter for the PS for comparison. We used the MSE $\left(\sum[(\widehat{PS} - PS)^2]/500\right)$ to measure estimation quality because it captures both bias and variance and is helpful for comparing the efficiency of estimators regardless of their bias (e.g. Davis-Stober, Dana, & Rouder, 2018). To test the inference for the frequentist methods, we reported the empirical coverage rate (ECR) of the 95% confidence intervals. The ECR is an estimate, so we computed Jeffrey's 90% CI (Brown, Cai, & DasGupta, 2001) to get a sense of the uncertainty in the ECR estimate.

We simulated data that we deemed somewhat realistic. The data generation process for binary assignment was $\mathbf{x} \sim \text{Bern}(q)$; where $q \in \{.3,\ .5\}$. For the outcome, we randomly sampled integers, $i = 0, 1, 2, \ldots, 100$ with replacement, according to their respective sampling weights, $f(i/100 \mid \alpha,\ \beta)$, where $f(\cdot)$ is the beta density function. For $\mathbf{x} = 0$ (the control group), $\alpha = 11/2,\ \beta = 11/6$; hence the control mean, standard deviation and skew were 75, 15 and $-0.71$ respectively. For $\mathbf{x} = 1$ (the treatment group), $\alpha = 12,\ \beta = 3$; hence the

treatment mean, standard deviation and skew were 80, 10 and $-0.71$ respectively. The treatment improved the outcome, and the treatment group was relatively homogenous, see Figure 2. The true PS was: $\left(\int_0^1 f(x \mid 11/2, 11/6)[1 - F(x \mid 12, 3)]dx\right)/100 = 59.63\%$, where $x = i/100$ and $F(\cdot)$ is the beta distribution function (equation 5 in Chechile, 2019). The data are supposed to represent test scores when there is a ceiling on performance, and the ability of test takers is not much lower than the ceiling on average. Thus, the data were negatively skewed. Additionally, the data were discrete (integers only) as is typical for tests, hence, there were ties in the data.

## Results

Across design conditions, the maximum $\widehat{R}$ and minimum $n_{\mathsf{eff}}$ were 1.002 and 3202 respectively, suggesting parameter estimation converged. We report the results for the study in Figure 3. Across both conditions, the rankits and nonparametric estimators were least biased, while the Bayesian and parametric estimators were downwardly and upwardly biased respectively. However, the Bayesian estimator had low variance such that the Bayesian estimator MSEs were relatively low, about 65% of the next best estimator: the rankits estimator. The MSE of the nonparametric estimator was always highest and the MSE of each estimator was always lower when the design was balanced. Finally, most methods had acceptable ECRs.

## Discussion

In this paper, we have presented a Bayesian estimator for the PS and demonstrated the adequacy of this estimator relative to other methods for estimating the PS at a small but common sample size in educational interventions. We consider the results from the simulation study to be informative for empirical practice. The non-parametric estimator by

Vargha and Delaney (2000) seemed unbiased but was high variance, such that the MSE was high. The parametric estimator by McGraw and Wong (1992) was upwardly biased, without substantive variance reduction. And the Bayesian estimator was downwardly biased with substantive gains in variance reduction, such that the Bayesian estimator was the most adequate estimator. Moreover, the Bayesian analysis allows for asking substantively interesting of the data. For example, one can ask: what is the probability that from 10 comparisons of students in the treatment and control group, the students in the treatment group will have a higher score at least 6 times, i.e. $\Pr(\text{PS} \geq 60\%)$?

Based on the limited simulation work in this study and the wealth of information a Bayesian analysis can provide (Kruschke, 2013), we would recommend Bayesian estimation of the PS. In the future, we will study the performance of the Bayesian approach under a larger variety of simulation conditions. And we will develop extensions for multilevel data structures, given that such structures are commonplace in educational settings (Goldstein, 1995).

## References

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. doi:10.1111/j.2044-8317.1978.tb00581.x

Brooks, M. E., Dalal, D. K., & Nolan, K. P. (2014). Are common language effect sizes easier to understand than traditional effect sizes? *Journal of Applied Psychology*, *99*(2), 332–340. doi:10.1037/a0034745

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16*(2), 101–133. doi:10.1214/ss/1009213286

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). doi:10.18637/jss.v076.i01

Chechile, R. A. (2019). A Bayesian analysis for the Mann-Whitney statistic. *Communications in Statistics - Theory and Methods*, 1–27. doi:10.1080/03610926.2018.1549247

Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed., p. 584). Wiley.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, *35*(3), 124–129. doi:10.2307/2683975

Davis-Stober, C. P., Dana, J., & Rouder, J. N. (2018). Estimation accuracy in the psychological sciences. *PLOS ONE*, *13*(11), e0207239. doi:10.1371/journal.pone.0207239

Goldstein, H. (1995). *Multilevel statistical models* (p. 384). London: Wiley.

Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, *1*(3), 403–420. doi:10.1214/06-BA116

Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *International Journal of Epidemiology*, *35*(3), 765–775. doi:10.1093/ije/dyi312

Hoffman, M. D., & Gelman, A. (2014). The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623. Retrieved from https://dl.acm.org/citation.cfm?id=2627435.2638586

Kraft, M. A. (2018). *Interpreting effect sizes of education interventions*. Brown University.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. doi:10.1037/a0029146

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361–365. doi:10.1037/0033-2909.111.2.361

Penfield, D. A., & McSweeney, M. (1968). The normal scores test for the two-sample problem. *Psychological Bulletin*, *69*(3), 183–191. doi:10.1037/h0025473

Reiczigel, J., Zakariás, I., & Rózsa, L. (2005). A bootstrap test of stochastic equality of two populations. *The American Statistician*, *59*(2), 156–161. doi:10.1198/000313005X23526

Ruscio, J., & Mullen, T. (2012). Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. *Multivariate Behavioral Research*, *47*(2), 201–223. doi:10.1080/00273171.2012.658329

Slavin, R., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis*, *31*(4),

500–506. doi:10.3102/0162373709352369

Soloman, S. R., & Sawilowsky, S. S. (2009). Impact of rank-based normalizing transformations on the accuracy of test scores. *Journal of Modern Applied Statistical Methods*, *8*(2). doi:10.22237/jmasm/1257034080

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, *25*(2), 101–132. doi:10.3102/10769986025002101

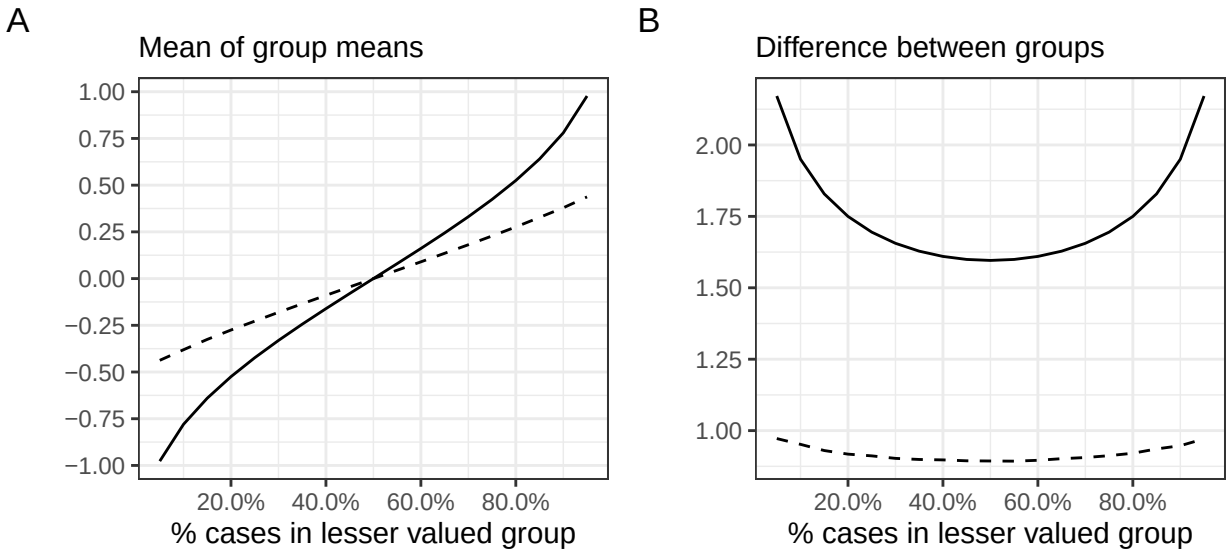


*Figure 1*. Mean of group means (panel A) and difference between groups (panel B) on the rankit scale as a function of proportion of cases in the lesser valued group, when there are no ties in the data. Solid lines represent theoretical maximums, dashed lines are empirical results when Cohen's d = 1 for simulated iid data of sample size of one million. Hence, Cohen's d of 1 translates to difference between group means that is just under one on the rankit scale.

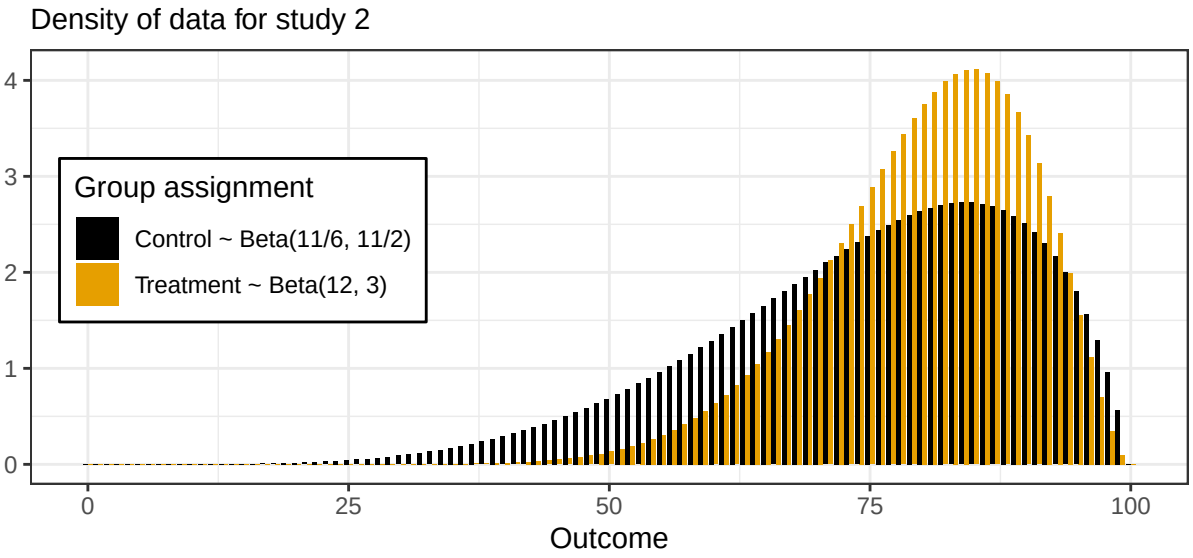


*Figure 2*. Density of data for simulation study. There were ties in the data, and each bar in the histogram represents an integer between 0 and 100.

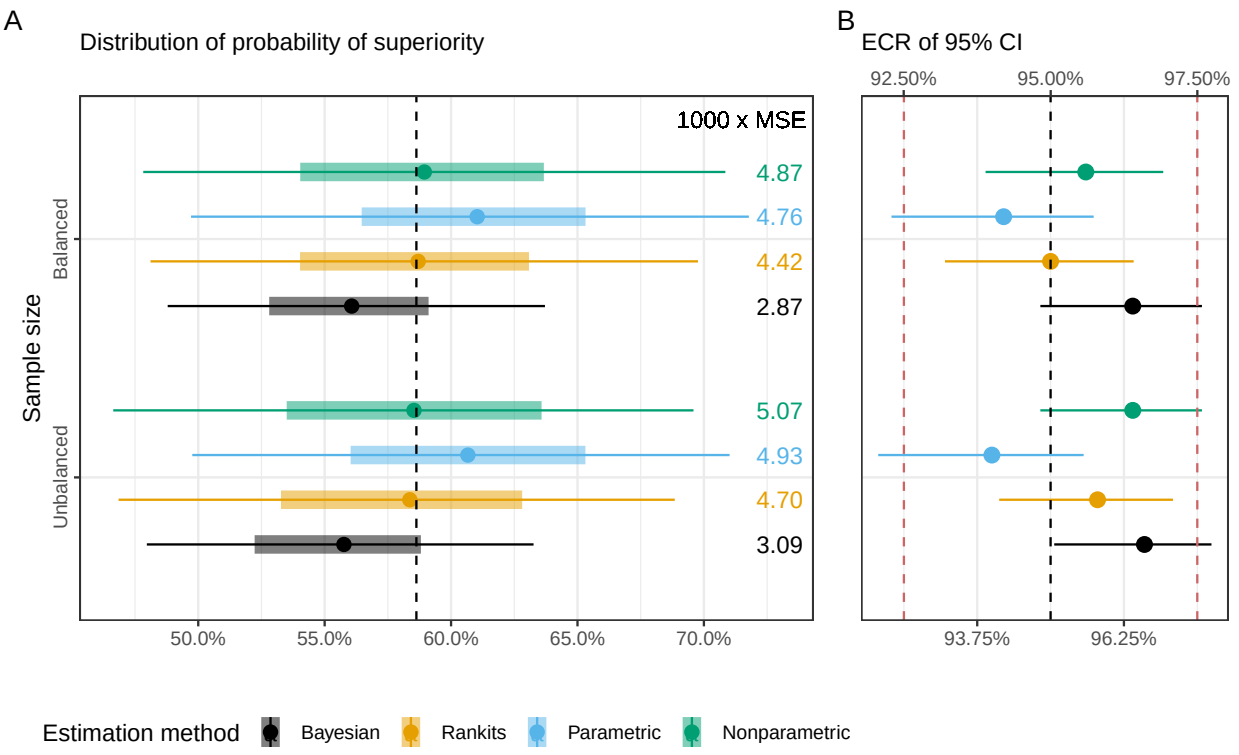


*Figure 3*. Results from simulation study 2. MSE values on the right of panel A are scaled by 1000 for convenience. The lineranges in panel A represent the 5th and 95th percentiles of the PS estimates, the shorter thicker bars represent the interquartile range, and the points are the mean PS estimates. Error bars around ECRs are Jeffrey's 90% CI. The black vertical dashed lines represent the true population parameter in panel A, and the nominal coverage rate in panel B. The red vertical dashed lines represent subjective limits within which acceptable ECRs should fall (liberal standard in Bradley, 1978).