Probability of superiority for comparing two groups of clusters

James Uanhoro
Research, Measurement & Statistics, Department of Educational Psychology
University of North Texas

Abstract

The probability of superiority (PS) has been recommended as a simple to interpret effect size for comparing two independent samples – there are several methods for computing the PS for this particular study design. However, educational and psychological interventions increasingly occur in clustered data contexts; and a review of the literature returned only one method for computing the PS in such contexts. In this paper, we propose a method for estimating the PS in clustered data contexts. Specifically, the proposal addresses study designs that compare two groups and group membership is determined at the cluster level. A cluster may be: (i) a group of cases with each case measured once, or (ii) a single case with each case measured multiple times, resulting in longitudinal data. The proposal relies on non-parametric point estimates of the PS coupled with cluster-robust variance estimation, such that the proposed approach should remain adequate regardless of the distribution of the response data. Using Monte Carlo simulation, we show the approach to be unbiased for continuous and binary outcomes, while maintaining adequate frequentist properties. Moreover, our proposal performs better than the single extant method we found in the literature. The proposal is simple to implement in commonplace statistical software and we provide accompanying R code. Hence, it is our hope that the method we present helps applied researchers better estimate group differences when comparing two groups and group membership is determined at the cluster level.

According to J. Cohen (1990), the “primary product of a research inquiry is one or more measures of effect size, not *p* values.” This is a view buttressed by Wilkinson and the APA Task Force on Statistical Inference (1999), who advised that researchers should “always present effect sizes for primary outcomes” (p. 599). Given that effect sizes are sample statistics, they are only estimates, thus, it is wise to communicate the uncertainty about such estimates. Confidence intervals are a useful tool for communicating uncertainty; the publication manual of the American Psychological Association (2020) describes the reporting of confidence intervals as “generally the best reporting

The Version of Record of this manuscript is published in Behavior Research Methods, and is available online at https://doi.org/10.3758/s13428-022-01815-6.

strategy" (p. 88).

We focus on the probability of superiority (PS) as an effect size for comparing two samples. 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) used the term *common language effect size* to describe what we call the PS. The PS as a term was introduced by Grissom (1994); other names for this measure include: measure of stochastic superiority (Klotz, 1966), stress-strength relation (Kotz, Lumelskii, & Pensky, 2003), *A* (Vargha & Delaney, 2000) and several others.

In this paper, we extend the PS to two-level clustered data contexts. Specifically, we propose a method for estimating the PS that would be adequate for two-arm cluster randomized trials (CRT), or when the grouping variable or experimental assignment is at level two. Our survey of the literature uncovered one approach for estimating the PS in such contexts (Zou, 2021). We propose an alternative method and compare the methods using Monte Carlo simulation.

In the remainder of this introduction section, we motivate the need for estimating the PS in clustered data, we review the extant method for estimating the PS in Zou (2021), then we present our proposal. In the next section, we run Monte Carlo simulations to assess the adequacy of our proposed method and compare its performance to the method in Zou (2021). Afterwards, we provide two applied examples for demonstration. We conclude with a discussion section where we address potential extensions to the method we have proposed.

Why estimate the PS?

Kelley and Preacher (2012) defined effect size as "a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest." This broad definition leaves room for different conceptualizations of effect size. The most common conceptualization is differences in location between two groups with resulting effect sizes such as the simple or standardized mean difference.

The PS relies on a different conceptualization of effect size to location differences. The PS can be motivated as an ordinal effect size (Cliff, 1993) that is often more aligned with the hypotheses of experimenters i.e. the chance that values from one group exceed values from a second group. This rationale is particularly important when the interval property of response data is not guaranteed. In such situations, the usefulness of mean differences is questionable. Even when the interval property of data is guaranteed, differences in means may be non-representative of the bulk of the data (e.g. under severe skew), or non-robust (e.g. under data contamination). The PS may be then be alternatively motivated as a robust measure for two-group comparison (e.g. Cliff, 1993; Li, 2015; Vargha & Delaney, 2000).

Another motivation for the PS is its ease of interpretation. McGraw and Wong (1992) called the PS the common language effect size because they expected that nonstatisticians will easily understand the term. Moreover there is empirical evidence to support the view that the PS is easier to interpret than other commonplace effect sizes (Brooks, Dalal, & Nolan, 2014). This motivation and the supporting evidence suggest that the PS should be computed for data whether or not they are interval.

To demonstrate the utility of the PS, consider a hypothetical comparison of two groups A and B on an outcome with the following effect sizes: mean difference (A − B) of 5, a standardized mean difference of 0.2, and a PS (A > B) of 55%. Were we to meet an individual from group A and another from group B, there is little we can conclude about the differences between both individuals on the basis of the location differences without additional information about (the distribution of) the response data in the original study. However, we can guess that there is an 11–in–20 chance that the individual from group A has a higher score than the individual in group B.

Finally, we note that methods for estimation of and inference about the PS for comparing two independent samples have reached a mature stage. Vargha and Delaney (2000) provides a nonparametric approach for inference that is adequate regardless of the distribution of the observed data. Moreover, Kotz et al. (2003) contains several parametric formulae for the PS that can improve estimation efficiency when the parametric model is adequate for the data. Additionally, results from Ruscio and Mullen (2012) show that the bias-corrected and accelerated (BCa) bootstrap maintains very well-behaved confidence intervals for the PS amongst a variety of analytic and bootstrap-based options. Beyond comparison of two independent samples, there is work by Li and Waisman (2019) and Li and Tze (2021) that advance estimation and inference about the PS for bivariate relationships. Ruscio and Gera (2013) cover extensions of the PS to paired samples and multi-group settings. Probabilistic index models (De Schryver & De Neve, 2019; Thas, Neve, Clement, & Ottoy, 2012) represent a regression framework for PS estimation and inference that allows for

PS estimation under a variety of settings including covariate adjustment.

Clustered data are increasingly present in educational and psychological research (McNeish, Stapleton, & Silverman, 2017). However, methods for estimating the PS in these contexts are sparse – in our review of the literature, we found only one method. This is in contrast to standardized mean differences where there is substantial work on estimation and inference for a variety of clustered data study designs (e.g. Hedges, 2011, 2016; Lai & Kwok, 2014, 2016). Given the increasing prevalence of these designs and the benefits of PS estimation, there is need to extend the PS to clustered data contexts.

Estimating the PS in two-level clustered data contexts

We begin this section by reviewing the single approach we found in the literature for estimating and performing inference about the PS when group membership is at level two.

Extant method based on placement scores. Zou (2021) developed an approach that begins by calculating placement scores (Delong, Delong, & Clarke-Pearson, 1988).[1] Given two groups A and B, the placement score for each case i in group A is the percentile of case i's response data point within group B response data. The reverse is done for each case in group B. One then regresses the placement scores on a group indicator with a random intercept on cluster membership. The group difference coefficient and its standard error is used to compute confidence intervals (CI) based on the t distribution with $G - 2$ degrees of freedom, where G is the number of clusters. A simple linear transformation transforms the group difference coefficient and its CI to the PS and its CI, $\text{PS} = (\text{coef} + 1)/2$. Zou (2021) explored some alternative methods for computing confidence intervals, and found the applying an ArcSinh transformation during interval estimation can improve inference especially when the PS is low or high.

Current proposal

Our proposal for estimation and inference about the PS begins with the nonparametric formulation by Vargha and Delaney (2000) for two independent samples:

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

[1] While the method is based on the work of Delong et al. (1988), we believe Hanley and Hajian-Tilaki (1997) are responsible for coining the term “placement”.

where $\#(\cdot)$ is the count function, $\mathbf{y}_1$ and $\mathbf{y}_2$ are the data for group 1 and group 2 respectively, with n_1 and n_2 as the corresponding sample sizes. 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. Under this formulation, the PS is directly proportional to the Mann-Whitney U-statistic (Vargha & Delaney, 2000).

We note that one can perform inference about $\widehat{\text{PS}}$ by treating the task as inference about a proportion with a given sample size $(n_1 + n_2)$. There are several methods for inference about such a proportion (Newcombe, 1998), but we have several desiderata. We desire a method that (i) can easily be extended to multiple proportions; (ii) does not introduce parametric assumptions; and (iii) has potential to account for covariates such that the approach can be scaled up. Requirement (i) includes most methods for inference about proportions. Requirement (ii) eliminates methods with important distributional assumptions such as beta-regression, while requirement (iii) suggests the need for a regression style model.

The fractional regression model of Papke and Wooldridge (1996), a quasi-likelihood approach, meets all three requirements. This is a model for probabilities with 0 and 1 as plausible outcome values, as opposed to most probability models where 0 and 1 are not plausible values. There is no need to specify the correct distribution for the probabilities to obtain consistent estimates. However, heteroskedasticity consistent standard errors are needed for adequate inference.

With this background, we now present the method to address the scenario where group membership occurs by cluster. Consider the hypothetical scenario where 10 clusters were assigned to the treatment group, and 20 clusters were assigned to the control group. To estimate the PS in this scenario, we recommend the following steps:

1. Calculate pairwise PS estimates using equation 1 comparing each cluster in the treatment group to all clusters in the control group. There will be 200 (10×20) PS estimates.

2. Perform intercept-only weighted fractional regression with the PS estimates as the outcome, and the weights as the total number of cases for each comparison. A weighted analysis ensures clusters with larger number of cases contribute more to coefficient estimation. The data for the regression will comprise 200 PS estimates and 200 sample sizes. One can assume any reasonable link function for proportion data (e.g. logit, probit) – we assume logit in this

paper.[2]

3. The intercept can then be transformed to the original probability scale to obtain the PS.

For each cluster j in the treatment group and each cluster k in the control group, the pairwise PS estimates in step 1 are:

$$\widehat{PS}_{jk} = [\#(\mathbf{y}_j > \mathbf{y}_k) + 0.5 \times \#(\mathbf{y}_j = \mathbf{y}_k)]/(n_j n_k) \quad \text{for } j \in \{1, \ldots, J\}, k \in \{1, \ldots, K\} \tag{2}$$

where $\mathbf{y}_.$ and $n_.$ are the data and sample sizes for their corresponding groups.

The logistic regression in step 2 maximizes the log-likelihood:

$$\sum_{j=1}^{J}\sum_{k=1}^{K}(n_j + n_k)\Big[\widehat{PS}_{jk} \ln\big[G(\hat{\theta})\big] + (1 - \widehat{PS}_{jk}) \ln\big[1 - G(\hat{\theta})\big]\Big] \tag{3}$$

where $\hat{\theta}$ is the intercept on the logit-scale, and $G(\hat{\theta}) = (1+e^{-\hat{\theta}})^{-1}$, the inverse-logit function. Hence, $G(\hat{\theta})$ is the estimated PS. Given that this is an intercept-only logistic model, $G(\hat{\theta})$ is simply the weighted mean of pairwise PS estimates:

$$\frac{\sum_{j=1}^{J}\sum_{k=1}^{K}(n_j + n_k)\widehat{PS}_{jk}}{\sum_{j=1}^{J}\sum_{k=1}^{K}(n_j + n_k)} \tag{4}$$

One problem with the above approach is that the outcome data for the fractional regression (pairwise PS estimates) are not independent from each other. Each PS estimate comes from two clusters, and the clusters repeat in the data. Hence, the resulting standard errors underestimate coefficient variability. We apply cluster-robust inference to remedy this:

1. Apply a two-way cluster-robust variance estimator (CRVE, Cameron & Miller, 2015) using the two cluster IDs that form each PS.

2. Compute the CI for the intercept on the logit-scale using the updated intercept standard error. The interval should be computed using the t distribution with $J + K - 2$ degrees of

[2] In R, the regression would be: glm(formula = PS ∼ 1, family = quasibinomial, weights = N), where PS are the 200 $\widehat{\text{PS}}$ estimates and N are the 200 sample sizes.

freedom rather than the normal distribution – this can improve inference especially when the number of clusters is small (Cameron & Miller, 2015; Donald & Lang, 2007).

3. Transform the interval back to the probability scale to obtain the CI for the estimated PS.

Assuming $se(\hat{\theta})$ is the resulting cluster-robust standard error for the intercept, the two-tailed CI for the PS is: $G\left(\hat{\theta} \pm se(\hat{\theta}) \times t_{1-\alpha/2,\nu}\right)$, where $\nu = J + K - 2$. In summary, the approach we present is a method for inference about the weighted mean of PS estimates obtained from pairwise comparisons of clusters.

In the next section, we run Monte Carlo simulations to assess the adequacy of our proposed method and compare its performance to the method in Zou (2021).

Simulation studies

In this section, we run three simulation studies to test the adequacy of the proposed fractional regression approach with CRVE. The context for the first study was a two-arm cluster randomized trial with a normally distributed continuous outcome and a null effect. We compared the fractional regression approach to the BCa intervals – the optimal method for independent samples – which ignore clustering and to the method based on placement scores (Zou, 2021). Both the fractional regression and placement score approaches performed favourably, while the BCa intervals are shown to be inadequate – this is to be expected since the BCa intervals assume the data are independent. In the second study, we stress-tested both the fractional regression and placement score approaches with non-normal continuous outcomes to see the conditions under which they are likely to fail. The fractional regression approach was adequate across all conditions, while the method based on placement scores was not. In the final study, we tested the adequacy of both approaches applied to a binary outcome – both methods were adequate. The specific placement score approach we used was multilevel regression on the placement scores with the ArcSinh transformation for interval estimation – the most efficient option based on results in Zou (2021). The specific CRVE we used was the CR3 variant (Bell & McCaffrey, 2002).

Assessment metrics. For all simulations, we assessed the approaches for accuracy and inference. We defined accuracy using the mean-squared error (MSE), as it subsumes both bias and variability of the estimator. We note ahead of time that all approaches were unbiased across all conditions such that the MSE was a measure of estimator variability. We defined adequate inference by the ability of the 95% CI to maintain a 95% empirical coverage rate (ECR). We adopted the liberal condition in Bradley (1978), such that we assumed the ECR adequate whenever the estimated ECR fell between 92.5% and 97.5%.

All normal distribution notation follows mean-variance notation. All simulation code is available at `https://osf.io/xd3ba/`.

Simulation study 1

In this study, we simulated data according to the following model:

$$x_c \sim \text{Bernoulli}(p), \;\; \theta_c \sim \mathcal{N}(0, \tau), \quad y_i \sim \mathcal{N}(\theta_{c[i]}, 1 - \tau) \tag{5}$$

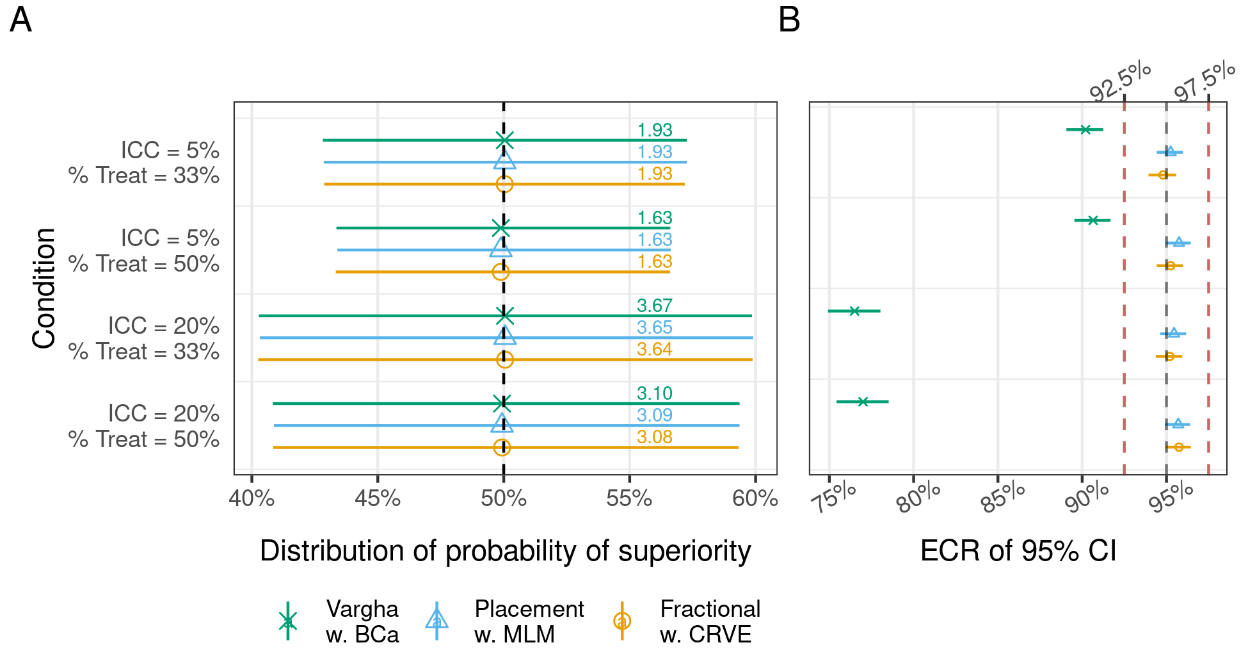
where x_c is a treatment indicator (1 = treat, 0 = control) for cluster c – it is Bernoulli with mean p. θ_c is the random intercept which varies by cluster membership – it is normal with mean 0 and variance τ. And y_i is the outcome for person i in cluster c – it is normal with a mean that depends only on the random intercept, and has error variance, $1 - \tau$. The treatment assignment has no influence on the outcome, hence this scenario is an exact null effect for the treatment and the population PS was 50%. Additionally, the model above functions such that τ is the intra-class correlation (as long as $\tau < 1$).

Simulation conditions. We assumed there were 30 clusters, with an average of 10 cases per cluster,[3] resulting in a total sample size of 300 cases on average. A minimum of 30 clusters is a common rule of thumb for clustered data (Huang, 2018), while 10 mimics the size of small classrooms. We created four conditions by varying both p ($p \in \{33.\bar{3}\%, 50\%\}$) and τ ($\tau \in \{5\%, 20\%\}$). Our expectation is that when p is lower than .5 (signifying unbalanced assignment to treatment and control), the estimated effect would be more variable. We similarly expected the estimated effect to be more variable with higher ICCs, as this corresponds to greater noise at the level of the treatment variable. Additionally, we expect the BCa interval to fail more at higher ICCs given the increased lack of case independence. We ran 2,000 replications per condition using the SimDesign package (Chalmers & Adkins, 2020) in R.

[3] The exact number of cases per cluster was round $(\mathcal{N}(10, 1))$ to simulate some variability into the process.

Figure 1

Results from simulation study 1



Note. Panel A. Numbers are MSE values, multiplied by 1000 for convenience. The points are the mean PS, errorbars represent the 5th and 95th percentiles of the PS estimates across replications. The black vertical dashed line represents the true population PS of 50%. Panel B. Errorbars around ECRs (points) are Jeffrey's 90% CI. The black vertical dashed line represents the nominal coverage rate of 95%. The outer red vertical dashed lines represent subjective limits within which acceptable ECRs should fall (liberal standard in Bradley (1978)).

Results. We report the results in Figure 1. All approaches resulted in unbiased estimates of the PS and comparable estimation efficiency within any given condition. As expected, estimates were more variable with unbalanced assignment of clusters to the groups and when the ICC was higher. Notably, the fractional regression and placement score approaches maintained adequate inference across the four conditions. The BCa intervals failed for all conditions with increased failure at higher ICCs.

These results rule out the BCa intervals as an inferential tool for the PS when the grouping variable is at level two. We now stress-test the two other approaches to see the conditions under which the approaches are likely to fail.

Simulation study 2

In this study, we tested the clustered data approaches (fractional regression and placement scores) under the assumption that the data were bounded continuous data between 0 and 1. Such data would be commonplace in educational applications, most commonly as grades, or psychological applications, most commonly as the average of several items with Likert response scales or visual analog scales, rescaled to the 0–1 interval, or whenever data are analyzed as percentage of maximum possible scores (P. Cohen, Cohen, Aiken, & West, 1999). We assumed the data were logit-normal i.e. the logit of the data were normally distributed and simulated data according to the following model:

$$
\begin{aligned}
x_c &\sim \text{Bernoulli}(p),\ \theta_c \sim \mathcal{N}\left(\alpha + \beta \times x_c, \tau \times (\delta + \gamma \times x_c)\right) \\
y_i &\sim \text{Logit-normal}\left(\theta_{c[i]}, (1-\tau) \times (\delta + \gamma \times x_{c[i]})\right)
\end{aligned}
$$

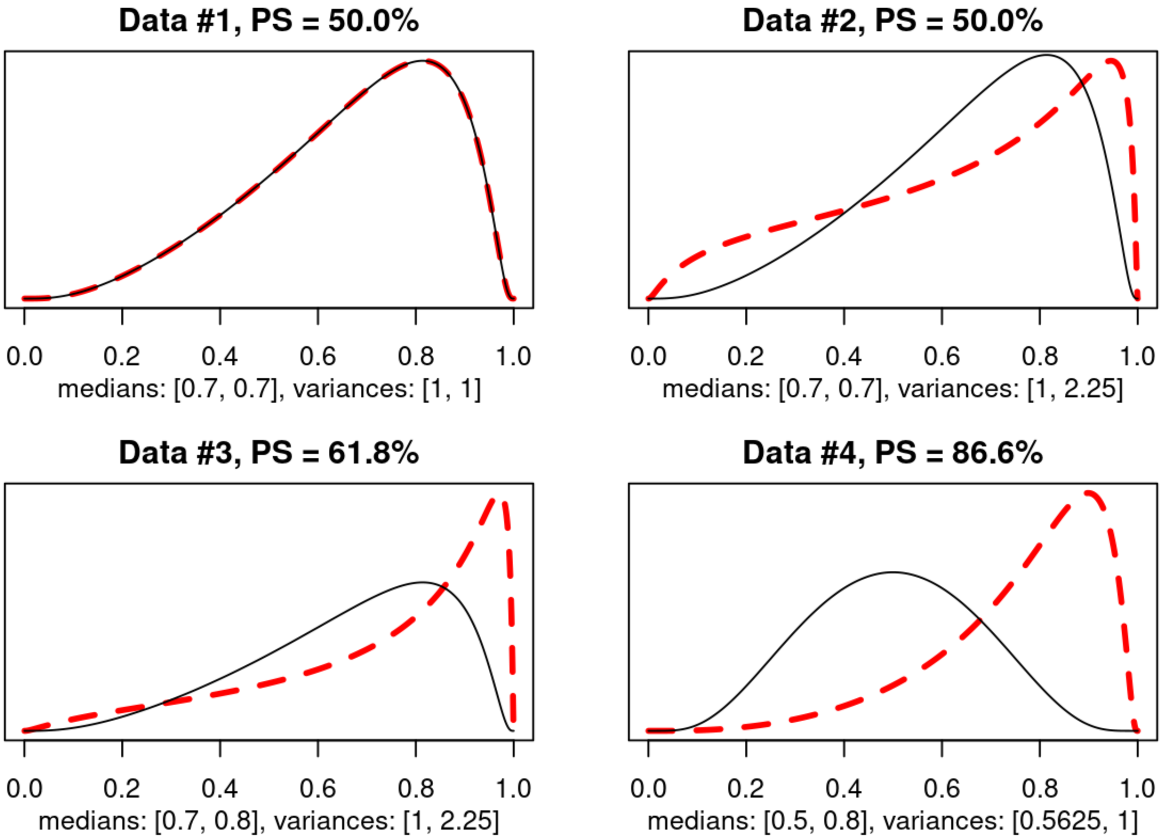
where x_c reains an indicator for assignment to the treatment group. We permitted the two groups to have different means and variances on the normal scale, resulting in different shapes for the data. Additionally, for the logit-normal distribution, the mean of the normal data prior to logit-transformation is the median of the resulting bounded data since the median is invariant under monotonic transformations.

The PS is invariant under monotonic transformations like the logit, such that the PS formula for normal data continues to apply to the PS for the resulting proportions after the logit transformation. Hence the population PS was $\Phi(\beta/\sqrt{\delta + \delta + \gamma})$ (McGraw & Wong, 1992, where Φ is the standard normal distribution function).[4] Finally, we varied α, β, δ and γ to produce four distinct patterns for the data which we discuss below, also see Figure 2.

[4] Broadly, one can use integration to obtain the population PS given known parametric distributions for the data. In the case of continuous variables, the PS is: $P(X > Y) = P(Y = y \,\&\, X > y) = P(Y = y) \times P(X > y) = f_Y(y) \times (1 - F_X(y)) = \int_a^b f_Y(y)(1 - F_X(y))dy$ where $f_Y(\cdot)$ is the density function of Y, $F_X(\cdot)$ is the distribution function of X, and a and b are the known theoretical limits of X and Y.

Figure 2

Data patterns for simulation study 2



Note. Distributions of the control (solid black) and treatment (dashed red) groups on the 0–1 scale.

Data #1. These data represent a situation where the treatment and control groups are identical, i.e. PS = 50%. Both groups have left-skewed distributions (e.g. participants' scores were relatively high) with median score of 70%. Precisely: $[\alpha, \beta, \delta, \gamma] = [0.847, 0, 1, 0]$.

Data #2. These data represent a situation where the treatment increases the scores of some participants while causing others to fall behind resulting in PS of 50%. Both groups have left-skewed distributions with median score of 70%. Precisely: $[\alpha, \beta, \delta, \gamma] = [0.847, 0, 1, 1.25]$. In this scenario, the mean of the control group actually exceeds the mean of the treatment group.

Data #3. These data represent a situation where the treatment mostly increases the scores of participants resulting in PS of 61.8%. Both groups have left-skewed distributions with treatment and control median scores of 70% and 80% respectively. Precisely: $[\alpha, \beta, \delta, \gamma] = [0.847, 0.539, 1, 1.25]$.

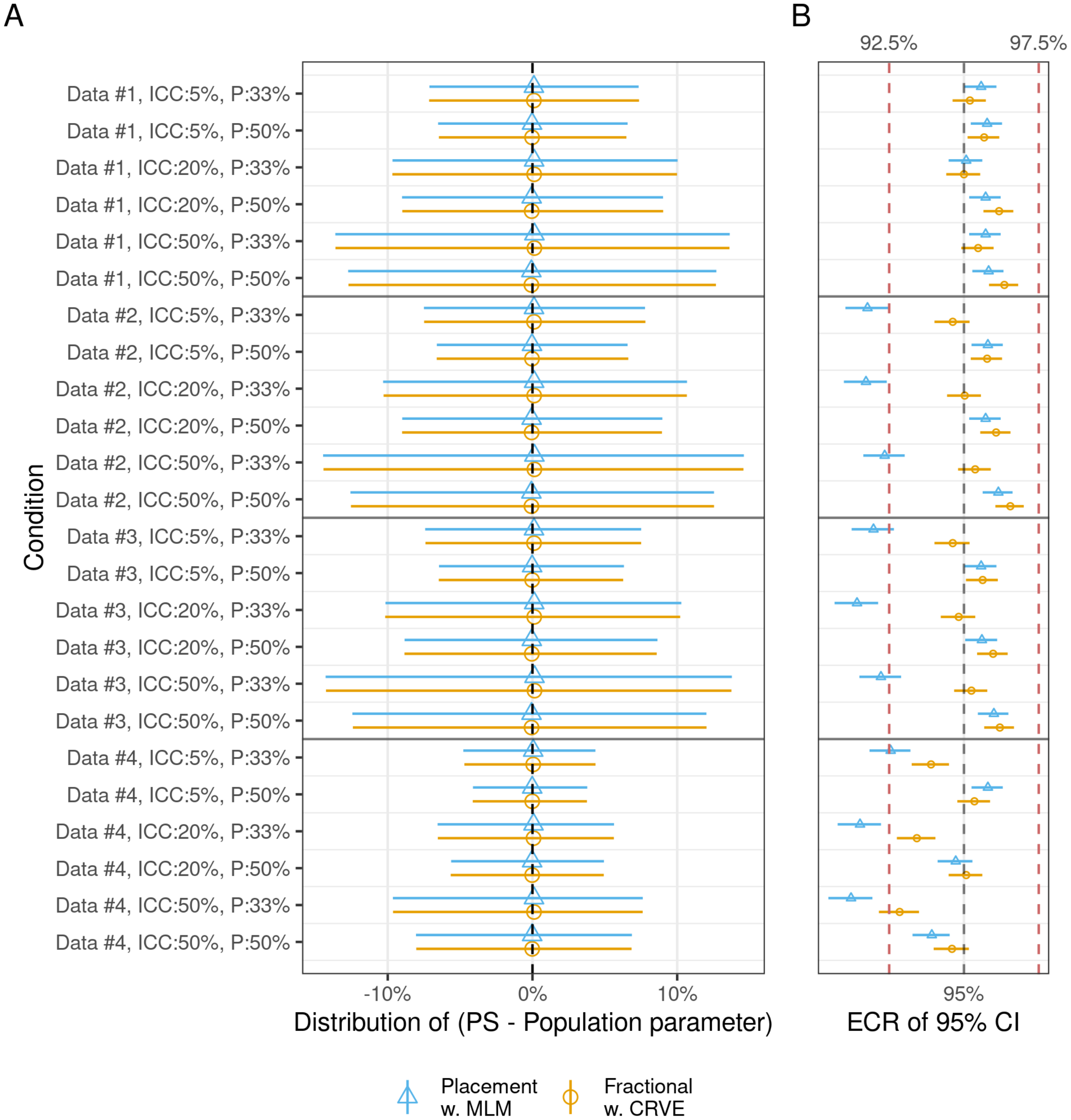
Data #4. These data represent a situation where the treatment overwhelmingly increases the scores of participants resulting in PS of 86.8%. Such a PS would be unrealistic for the bulk of educational interventions, though it might be possible with psychological data.[5] Inference about the PS is more difficult as the PS becomes more extreme (Ruscio & Mullen, 2012), hence we include this pattern as a more extreme test of both approaches. The control group is symmetric with median of 50% and the treatment group is left-skewed with median of 80%. Precisely: $[\alpha, \beta, \delta, \gamma] = [0, 1.386, 0.5625, 0.4375]$.

For each of the data patterns above, we retained the sample size specifications from study 1 (30 clusters, 10 cases per cluster on average), maintained $p \in \{33.\bar{3}\%, 50\%\}$ but altered $\tau \in \{5\%, 20\%, 50\%\}$ – we included a 50% ICC condition that would be realistic for longitudinal data where each cluster is a case measured multiple times. p and τ maintain their meanings from study 1. This resulted in six conditions per data pattern and 24 simulation conditions in all. We ran 4,000 replications per condition.

[5] Assuming the data are normally distributed, a PS of 86.8% is approximately a Cohen's d value of 1.6. This follows from the relation $d = \Phi^{-1}(\text{PS}) \times \sqrt{2}$.

Figure 3

Results from simulation study 2 – non-normal outcomes



Note. Panel A. The points are the bias and errorbars represent the 5th and 95th percentiles of the bias estimates. The black vertical dotted line represents the ideal bias of 0. Panel B. Errorbars around ECRs (points) are Jeffrey's 90% CI. The black vertical dashed line represents the nominal coverage rate of 95%. The outer red vertical dashed lines represent subjective limits within which acceptable ECRs should fall (liberal standard in Bradley (1978)).

Results. We report the results in Figure 3. Both approaches remained unbiased with comparable estimation efficiency within any given condition. The fractional regression approach maintained adequate inference across all 24 conditions, although its ECR was just within acceptable limits for the second to last row in Figure 3 with: (extreme) data pattern #4, unbalanced assignment of clusters to groups and ICC of 50%. On the other hand, outside of data pattern #1, the placement score approach was often inadequate with unbalanced assignment of clusters to the treatment group (P:33% conditions) – this condition was untested in most simulations in Zou (2021). These results show that there are conditions under which the fractional regression approach is more reliable than the approach based on placement scores.

Simulation study 3

In this study, we tested the clustered data approaches (fractional regression and placement scores) for binary outcome data. To simulate such data, we assumed the data were latent normal prior to dichotomization:

$$x_c \sim \text{Bernoulli}(p),\ \theta_c \sim \mathcal{N}\left(\alpha + \beta \times x_c, \tau\right),\ \eta_i \sim \mathcal{N}\left(\theta_{c[i]}, 1-\tau\right),\ y_i = \mathbf{I}(\eta_i > 0)$$

where x_c remains an indicator for treatment assignment. The above equation for the binary outcome (y_i) is equivalent to a probit regression formulation for the data. The population PS is $(\kappa + 1)/2$, where κ is the risk difference or average difference in probabilities between both groups ($\Phi(\alpha + \beta) - \Phi(\alpha)$, Zou, 2021). We varied α and β to obtain different means for each group:

1. Data #1: Both groups have a mean of 70% – a null condition with population PS of 50%. Precisely: $\alpha = \Phi^{-1}(.7),\ \beta = 0$.
2. Data #2: The control group has a mean of 70% and the treatment group has a mean of 80%, a non-null condition with population PS of 55%. Precisely: $\alpha = \Phi^{-1}(.7),\ \beta = (\Phi^{-1}(.8) - \Phi^{-1}(.7))$.
3. Data #3: The control group has a mean of 50% and the treatment group has a mean of 80%, a non-null condition with population PS of 65%. A 30%-point increase would be a marked

effect on a binary outcome. In this exact scenario, this effect amounts to an odds-ratio of 4, $\exp\left(\text{inv-logit}(.8) - \text{inv-logit}(.5)\right)$. Precisely: $\alpha = \Phi^{-1}(.5)$, $\beta = \left(\Phi^{-1}(.8) - \Phi^{-1}(.5)\right)$.

For each of the data patterns above, we retained the sample size specifications from study 2 (30 clusters, 10 cases per cluster on average), and maintained $p \in \{33.\bar{3}\%, 50\%\}$ and $\tau \in \{5\%, 20\%, 50\%\}$. p and τ maintain their meanings from the previous studies. This resulted in six conditions per data pattern and 18 simulation conditions in all. We ran 4,000 replications per condition.

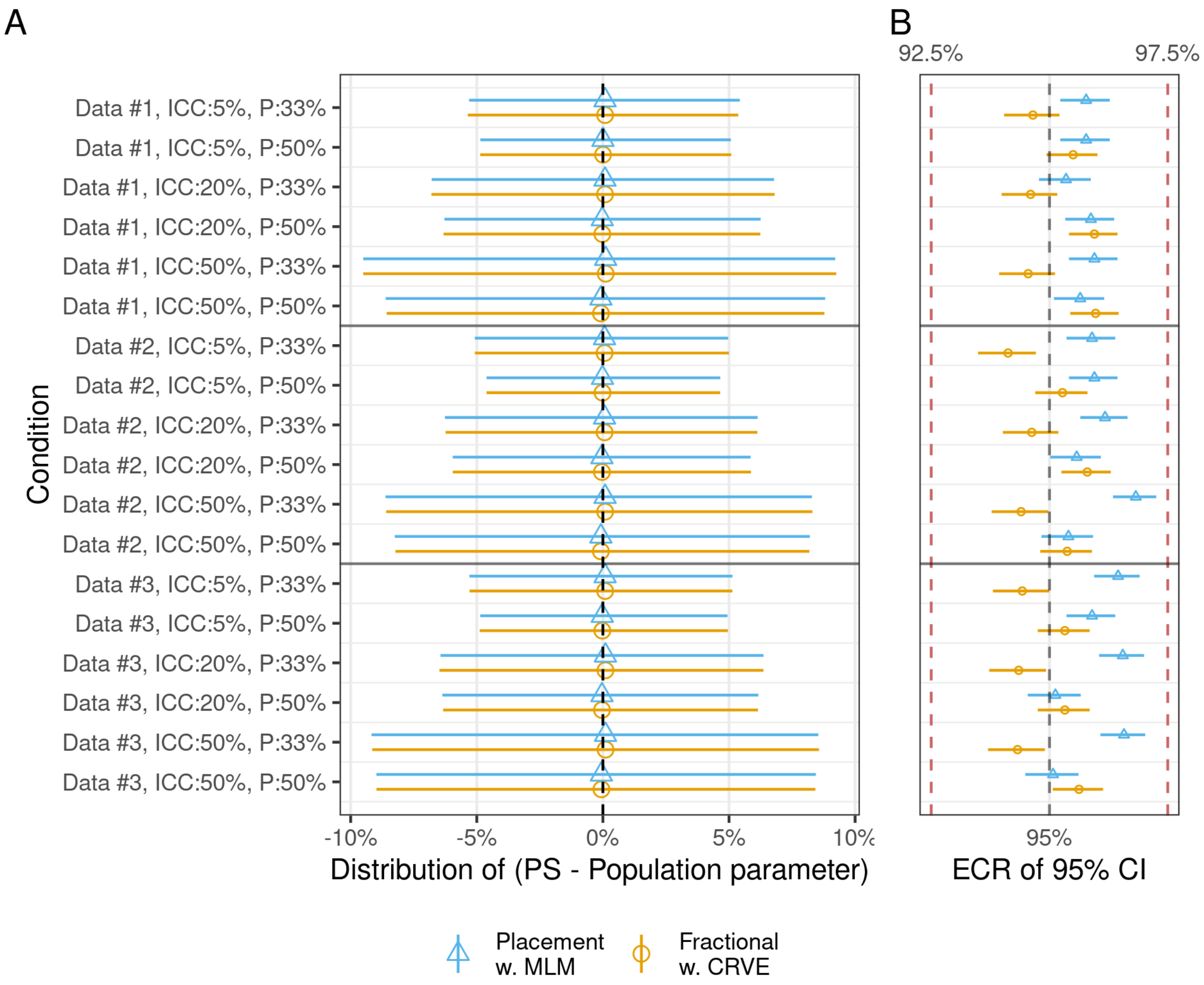
Results. We report the results in Figure 4. Both approaches remained unbiased with comparable estimation efficiency within any given condition and maintained adequate inference across all conditions. This result suggests both approaches may be relatively robust when applied to binary data. It is worth noting that under unbalanced cluster membership, the ECR for the placement scores approach was consistently higher than the ECR from the fractional regression approach.

Summary of simulation results

Fractional regression with CRVE can be used to estimate the PS while maintaining adequate inference. This approach should lead to similar conclusions as the placement scores approach in many scenarios. However, there may be disagreement between both methods when cluster assignment to or cluster membership in groups is unbalanced. When such divergence occurs, our findings lead us to believe the fractional regression approach with CRVE may be more reliable.

Figure 4

Results from simulation study 3 – binary outcomes



Note. Panel A. The points are the bias and errorbars represent the 5th and 95th percentiles of the bias estimates. The black vertical dotted line represents the ideal bias of 0. Panel B. Errorbars around ECRs (points) are Jeffrey's 90% CI. The black vertical dashed line represents the nominal coverage rate of 95%. The outer red vertical dashed lines represent subjective limits within which acceptable ECRs should fall (liberal standard in Bradley (1978)).

Data analysis examples

Code for the data analysis examples is available at `https://osf.io/xd3ba/`.

An example from a cluster randomized trial. Blair and Raver (2014) randomized schools to either a control or treatment group. Kindergartners in the treatment group were exposed to an approach “that embeds support for self-regulation, particularly executive functions, into literacy, mathematics, and science learning activities.” The researchers collected several outcomes. For this demonstration, we analyzed the pupils’ applied problems score at the end of first grade. After missing data deletion, there were 661 pupils in 31 schools (17 in the treatment group), about 21 pupils per school with the smallest school having 9 pupils.[6] The ICC for this outcome was 16%. The PS with BCa intervals (ignoring the clustering) was 53.6%, 95% BCa CI [49.5%, 58.2%]. The PS with 95% interval based on fractional regression with CRVE was 53.9%, 95% CI [43.2%, 64.2%]. And the results based on the placement scores was 53.8%, 95% CI [43.4%, 63.8%]. The intervals accounting for clustering had more than double the width of the BCa intervals. And though both intervals accounting for clustering were similar, fractional regression with CRVE intervals produced slightly wider intervals. Regardless of estimation method, there was about 54% chance that scores from pupils in the treatment group were higher than scores from pupils in the control group.

Describing existing differences between groups. We analyzed differences in math achievement between public and private school students in the high school & beyond dataset (Raudenbush & Bryk, 2002). There were 7,185 students from 160 schools (70 private schools), about 45 students per school with the smallest school having 9 students. The ICC for this outcome was 18%. We were interested in the probability that scores from private schools were higher than scores from public schools. The PS with 95% interval based on fractional regression with CRVE was 61.5%, 95% CI [58.0%, 64.9%]. Results based on the placement scores was 61.5%, 95% CI [57.9, 65.0%]. Again, these intervals were about the same, but the placement score intervals were slightly wider. There was about 61.5% chance that scores from private school students were higher than scores from public school students.

Discussion

We begin by reviewing the simulation results. The simulation showed that fractional regression with CRVE can be used to estimate the PS while maintaining adequate inference. We believe that this approach will lead to similar conclusions as the placement scores approach in many scenarios. However, when there is a disagreement between fractional regression and the placement score approaches, our findings lead us to believe the fractional regression approach with CRVE

[6] The paper mentions 29 schools, but the publicly available data has 31 school IDs.

may be more reliable. This improved performance of fractional regression is especially likely when cluster assignment to groups is unbalanced.

Additionally, we believe these findings to be robust to the distribution of the data. The fractional regression approach introduces no distributional assumptions for the PS estimates which are based on the nonparametric estimator of Vargha and Delaney (2000). The data for simulation study 2 were markedly non-normal, while simulation study 3 was based on binary data. Open questions about the proposed method with regard to inference include: How low can the number of clusters be? How high can the ICC be? How variable can the within-cluster sample sizes be? These questions are common to multilevel designs. We have shared our simulation code, making it easier for others interested in these questions to examine them.

We have laid out one approach to estimate the PS for a level-two grouping variable. We now speculate on some alternative methods that may be adequate. Cluster bootstrap methods (Field & Welsh, 2007) should maintain adequate inference for the PS computed from the original data. To do this, the PS would be calculated in the usual way (Vargha & Delaney, 2000) but the data would be resampled using cluster bootstrap methods to obtain inference about the estimated PS. Alternatively, one can take a Bayesian model-based approach that assumes the model for the data is rich enough to describe the distribution of the treatment and control groups. Given the assumed distribution, one can use either analytic methods or integration to compute the PS alongside the posterior samples of the PS. The disadvantage of this approach is that it is fully parametric hence the model must truly capture the data. However, when the model assumptions are met, we can expect the resulting PS estimates to be more efficient. Alternatively, one may modify the placement scores approach developed by Zou (2021) to use CRVE instead of multilevel modeling. We intend to investigate some of these approaches in the future.

An additional opportunity is the potential to adjust for covariates, with the intended benefit of reducing estimate uncertainty. This is one of the strengths of probabilistic index models, but these models have not been extended to clustered data contexts. If one only intends to adjust for level two covariates, probabilistic index models suggest one way forward for the proposed fractional regression approach. While computing the PS between each treatment cluster and all control clusters, one can compute the difference in the level-two covariates between the pair of clusters being compared and include the difference in covariates as predictors of the PS. This proposal has the potential to extend the fractional regression approach to account for covariates. We intend to explore this proposal in the future.

Finally, researchers in the behavioral sciences increasingly utilize more complex designs than two-level hierarchical models. For the PS to gain widespread use in these additional contexts, it is essential to extend computation of the PS to more contexts. We believe the fractional regression approach can be extended to variety of contexts. However, the exact specifications for PS computation and requirements for adequate inference would vary by design. Hence, we do not lay out the different specifications and requirements here. We intend to explore some of these specifications in the future.

Declarations

Funding. No funds, grants, or other support was received. The authors have no relevant financial or non-financial interests to disclose.

Open practices statement. All code for simulation studies and data analyses are available at `https://osf.io/xd3ba/`.

References

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). Washington, DC..

Bell, R. M., & McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multistage samples. *Survey Methodology*, *28*(2), 169–182.

Blair, C., & Raver, C. C. (2014, nov). Closing the achievement gap through modification of neurocognitive and neuroendocrine function: Results from a cluster randomized controlled trial of an innovative approach to the education of children in kindergarten. *PLOS ONE*, *9*(11), e112393. doi: 10.1371/JOURNAL.PONE.0112393

Bradley, J. V. (1978, nov). 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

Cameron, A. C., & Miller, D. L. (2015, apr). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372. doi: 10.3368/jhr.50.2.317

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology*, *16*(4), 248–280. doi: 10.20982/tqmp.16.4.p248

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494–509. doi: 10.1037/0033-2909.114.3.494

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312. doi: 10.1037/0003-066X.45.12.1304

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research*, *34*(3), 315–346. doi: 10.1207/S15327906MBR3403_2

De Schryver, M., & De Neve, J. (2019). A tutorial on probabilistic index models: Regression models for the effect size P(Y1 < Y2). *Psychological Methods*, *24*(4), 403–418. doi: 10.1037/MET0000194

Delong, E. R., Delong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*(3), 837–845.

Donald, S. G., & Lang, K. (2007). Inference with difference-in-differences and other panel data. *The Review of Economics and Statistics*, *89*(2), 221–233.

Field, C. A., & Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(3), 369–390. doi: 10.1111/j.1467-9868.2007.00593.x

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology*, *79*(2), 314–316. doi: 10.1037/0021-9010.79.2.314

Hanley, J. A., & Hajian-Tilaki, K. O. (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: An update. *Academic Radiology*, *4*(1), 49–58. doi: 10.1016/S1076-6332(97)80161-4

Hedges, L. V. (2011, jun). Effect sizes in three-level cluster-randomized experiments. *Journal of Educational and Behavioral Statistics*, *36*(3), 346–380. doi: 10.3102/1076998610376617

Hedges, L. V. (2016, oct). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370. doi: 10.3102/1076998606298043

Huang, F. L. (2018). Multilevel modeling myths. *School Psychology Quarterly*, *33*(3), 492–499. doi: 10.1037/spq0000272

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137–152. doi: 10.1037/a0028086

Klotz, J. H. (1966, September). The Wilcoxon, Ties, and the Computer. *Journal of the American Statistical Association*, *61*(315), 772–787. doi: 10.1080/01621459.1966.10480904

Kotz, S., Lumelskii, Y., & Pensky, M. (2003). *The stress-strength model and its generalizations.* World Scientific. doi: 10.1142/5015

Lai, M. H. C., & Kwok, O.-M. (2014, aug). Standardized mean differences in two-level cross-classified random effects models. *Journal of Educational and Behavioral Statistics*, *39*(4), 282–302. doi: 10.3102/1076998614532950

Lai, M. H. C., & Kwok, O.-M. (2016). Estimating standardized effect sizes for two-and three-level partially nested data. *Multivariate Behavioral Research.* doi: 10.1080/00273171.2016.1231606

Li, J. C.-H. (2015, oct). Effect size measures in a two-independent-samples case with nonnormal and nonhomogeneous data. *Behavior Research Methods*, *48*(4), 1560–1574. doi: 10.3758/S13428-015-0667-Z

Li, J. C.-H., & Tze, V. M. C. (2021, mar). Analytic and bootstrap confidence intervals for the common-language effect size estimate. *Methodology*, *17*(1), 1–21. doi: 10.5964/METH.4495

Li, J. C.-H., & Waisman, R. M. (2019, February). Probability of bivariate superiority: A non-parametric common-language statistic for detecting bivariate relationships. *Behavior Research Methods*, *51*(1), 258–279. doi: 10.3758/s13428-018-1089-5

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

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the Unnecessary Ubiquity of Hierarchical Linear Modeling. *Psychological Methods*, *22*(1), 114–140. doi: 10.1037/met0000078

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine*, *17*(8), 857–872. doi: 10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E

Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *Journal of Applied Econometrics*, *11*(6), 619–632. doi: 10.1002/(SICI)1099-1255(199611)11:6

Raudenbush, S. W., & Bryk, A. (2002). *Hierarchical Linear Models* (2nd ed.). Thousand Oaks: Sage.

Ruscio, J., & Gera, B. L. (2013, March). Generalizations and Extensions of the Probability of Superiority Effect Size Estimator. *Multivariate Behavioral Research*, *48*(2), 208–219. doi: 10.1080/00273171.2012.738184

Ruscio, J., & Mullen, T. (2012, mar). 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

Thas, O., Neve, J. D., Clement, L., & Ottoy, J.-P. (2012, sep). Probabilistic index models. *Journal* *of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(4), 623–671. doi: 10.1111/J.1467-9868.2011.01020.X

Vargha, A., & Delaney, H. D. (2000, jan). 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

Wilkinson, L., & the APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. *American Psychologist*, *54*(8), 594–604. doi: 10.1037/0003-066X.54.8.594

Zou, G. (2021, jun). Confidence interval estimation for treatment effects in cluster randomization trials based on ranks. *Statistics in Medicine*, *40*(14), 3227–3250. doi: 10.1002/SIM.8918