Connecting simulation to circumstance: Updates

JoonHo Lee (jlee296@ua.edu)

August 4, 2021

1 Introduction

Based on Luke's previous document, this report attempts to connect the main simulation factors to what kinds of multisite trials would have those simulation factors.

Since this issue is critical for contextualizing both our simulation study and real data analysis results, we need to take extra steps to clarify the connection. If our simulation setup is not plausible based on the evidence from past multisite trials, we should consider restructuring our simulation with updated data-generating models.

2 The simulation factors for sampling errors

A vector of the first-stage measurement or sampling errors, $\widehat{se}_j^2$, is originally generated by the combination of two factors:

1. $I = \{0.1, 0.5, 0.9\}$: the average reliability of the ML estimates $\hat{\tau}_j$
2. $R = \{1, 5, 10\}$: the heterogeneity of the $\widehat{se}_j^2$'s across the N sites

The resulting vector of simulated $\widehat{se}_j^2$'s is denoted as $\hat{\text{E}}$. We are actually talking about *true* sampling errors, not *estimated* ones, because we are involved in the context of true data generation for the simulation study.

3 Reflecting the magnitude of *real-world* cross-site impact variation

The average reliability I determines how informative the first-stage ML estimates are on average. A large I value indicates less noisy, more informative observed ML estimates $\hat{\tau}_j$'s. In the paper, I is given by

$$I = \frac{\sigma^2}{\sigma^2 + \text{GM}(\widehat{\text{se}}_j^2)}$$

where σ^2 represents the genuine heterogeneity in true site-specific effects τ_j between sites (i.e., $\text{Var}(\tau_j)$, the cross-site impact variation). $\text{GM}(\widehat{\text{se}}_j^2)$ indicates the geometric mean of $\widehat{se}_j^2$, $\exp\left(\sum_{j=1}^{N} \ln(\widehat{se}_j^2)\right)$.

We can rewrite this equation in terms of $\text{GM}(\widehat{\text{se}}_j^2)$:

$$\text{GM}(\widehat{\text{se}}_j^2) = \sigma^2 \cdot \frac{(1 - I)}{I}$$

Now the average level of within-site sample variance is a function of σ^2 and I.

Luke's previous document suggested using *effect size units*, which are defined as a multiple of an individual-level standard deviation for the original outcome measure, $\text{SD}(Y)$. For example, an estimate of the cross-site

impact variation, $\widehat{\sigma}^2$, might be presented as 0.20^2 in effect size units. This means that the magnitude of the cross-site impact variation equals 0.20 standard deviations of the outcome measure in its original units. In this case, we would have $\text{Var}(Y) = 1$.

Weiss et al. (2017) documented that about 62% of their estimates from 16 large multisite trials represented moderate (0.15) to large (0.25) cross-site impact variation. Thus, it is reasonable to set $\sigma^2 = 0.20^2$ for a moderately large amount of cross-site impact variation.

In our previous simulation setup, however, we fixed the cross-site impact variation to one ($\sigma^2 = 1.0$) in all the true data-generating models for the prior distribution G. Then, we will have **25 times** larger average level of the first-stage sampling error $\text{GM}(\widehat{\text{se}}_j^2)$:

```
# Create a function to calculate GM(SE^2)
mean_SE2 <- function(sigma = 0.20, I = 0.1){
  GM_SE <- (sigma^2*(1 - I))/I
  return(GM_SE)
}

# Difference in GM(SE^2) when sigma = 0.2 vs. 1.0
(GM_SE_0.2 <- mean_SE2(sigma = 0.20, I = 0.1))
```

```
## [1] 0.36
```

```
(GM_SE_1.0 <- mean_SE2(sigma = 1.00, I = 0.1))
```

```
## [1] 9
```

```
GM_SE_1.0/GM_SE_0.2
```

```
## [1] 25
```

To reflect the magnitude of cross-site impact variation from past multisite trials (e.g., $\sigma^2 = 0.20^2$) while fixing $\sigma^2 = 1.0^2$, we need to set the variance of the outcome measure in its original scale to 25 ($\text{Var}(Y) = 25$) instead of $\text{Var}(Y) = 1$. If we assume that $\sigma^2 = 0.15$ in effect size units, $\text{Var}(Y)$ equals 44.4 times larger than 1 (i.e., $\text{Var}(Y) = 44.4$) given $\sigma^2 = 1.0$.

If we fix $Var(Y)$ to 1, σ^2 represents the magnitude of cross-site impact variation in effect size units. If σ^2 is fixed to 1, on the other hand, $Var(Y)$ will now represent the outcome variance in its original scale, but reflecting the assumed magnitude of σ^2.

4 Connecting sampling errors to site sizes

The reasoning above is based on the following design-based sampling error estimator (Miratrix, Weiss, and Henderson 2021):

$$\widehat{\text{se}}_j^2 = \left(\frac{1}{n_j \cdot p_j} + \frac{1}{n_j \cdot (1 - p_j)}\right) \cdot \text{Var}(Y) = \frac{\text{Var}(Y)}{n_j \cdot p_j \cdot (1 - p_j)}$$

where n_j is sample size at site j and p_j is the proportion of units treated in site j (labeled as "propensity score" by Rosenbaum and Rubin (1983)). $\text{Var}(Y)$ is the outcome variance across sites and experimental conditions, which is assumed to be constant. This expression can be viewed as Neyman's classic formula under assumed homoskedasticity.

In a multisite trial, p_j can vary across sites by design, or more often, due to unobservable site differences. But here, we further assume constant p across the sites for simplicity. Taking geometric means on both sides of the equation results in the following:

$$\text{GM}(\widehat{\text{se}}_j^2) = \frac{\text{Var}(Y)}{p \cdot (1-p)} \cdot \text{GM}\left(\frac{1}{n_j}\right) = \kappa \cdot \text{GM}(n_j)^{-1}$$

for some constant κ. Previously, $\text{GM}(\widehat{\text{se}}_j^2)$ was determined by I and the fixed unit variance of cross-site impacts ($\sigma^2 = 1$). Now $\text{GM}(\widehat{\text{se}}_j^2)$ is a function of two constants, $\text{Var}(Y)$ and p, and one variable, n_j. Then, we can obtain a vector of simulated $\widehat{se}_j^2$'s, $\hat{\text{E}}$, by directly controlling those three factors, particularly the site sizes. Our new data-generating functions will reflect this relationship.

5 Required mean site size to achieve a given level of information (I)

Our natural next step is to define a function to generate $\text{GM}(n_j)$. We can solve for the geometric mean of site sizes to get a formula of required site sample size to achieve a given level of information (I):

$$\begin{aligned}\text{GM}(n_j) &= \kappa \cdot \text{GM}(\widehat{\text{se}}_j^2)^{-1} \\ &= \frac{\text{Var}(Y)}{p \cdot (1-p)} \cdot \frac{I}{\sigma^2 \cdot (1-I)}\end{aligned}$$

This formula is essentially the same as the one presented in Luke's document. But I intend to explicitly include the variation in outcomes as one factor because, in our setting, $\text{Var}(Y)$ is designed to reflect the assumed level of cross-site impact variation when σ^2 is fixed to 1.0.

Then we can create a function to calculate the average level of site sample sizes implied by our simulation factors:

```
# Create a function to calculate the average site sizes
# implied by simulation factors
GM_site_size <- function(I = 0.5, p = 0.50,
                         sigma_ES = 0.20, sigma_fixed = 1.0){

  # Var(Y) = the ratio between 1) fixed sigma and 2) sigma in effect size units
  varY <- sigma_fixed^2/sigma_ES^2

  # Calculate the average level of site sizes implied by I
  kappa <- varY/(p*(1 - p))
  GM_se2_inv <- I/(sigma_fixed^2 * (1 - I))
  GM_nj <- kappa*GM_se2_inv
  return(GM_nj)
}
```

A key feature of this function is to specify the researcher's assumed level of cross-site impact variation in effect size units (`sigma_ES`) with a fixed quantity of the variation used to simulate G (`sigma_fixed`). The ratio between these two quantities yields the outcome variance, $\text{Var}(Y)$.

Given this function, we can calculate the implied mean site size by specifying `I`, `p`, `sigma_ES`, and `sigma_fixed`. Consider the following specifications:

```
# Create a table crossing design factors
design_factors <- list(
  I = seq(from = 0.1, to = 0.9, by = 0.1),
  p = c(0.5, 0.7),
  sigma_ES = c(0.15, 0.20),
  sigma_fixed = 1.0
)
```

```
params <- design_factors %>%
  cross_df()
```

The function for calculating the implied mean site size is then applied to the specified simulation factors:

```
# Calculate average site sizes according to design factors
GM_nj <- pmap(params, GM_site_size)

df_result <- params %>%
  mutate(mean_nj = GM_nj) %>%
  unnest(cols = mean_nj) %>%
  mutate(mean_nj = round(mean_nj, 1))
```

The output below reports the implied mean site size for each level of I when `sigma_ES` and `p` are set to 0.2 and 0.5, respectively.

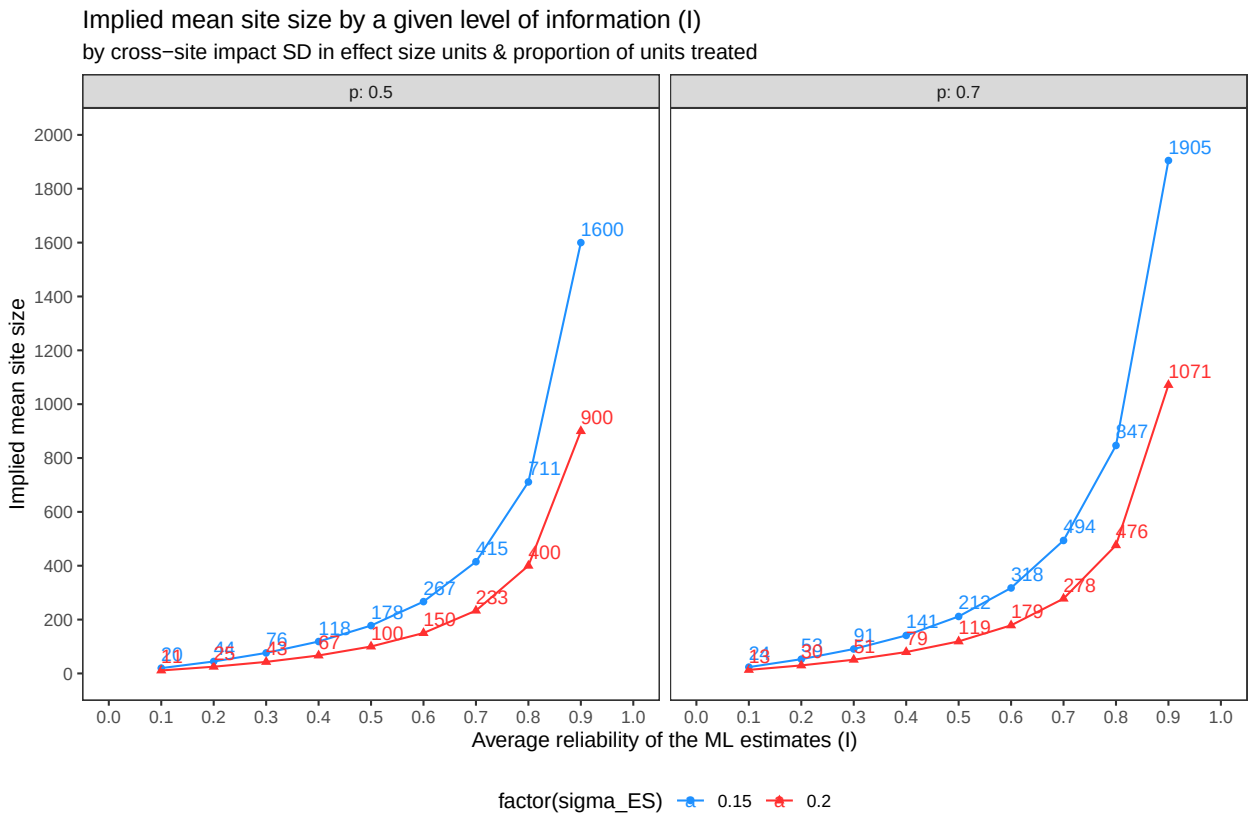
```
# Dispay results for a subset (p = 0.5, sigma_ES = 0.20)
df_result %>%
  filter(p == 0.5, sigma_ES == 0.20)
```

```
## # A tibble: 9 x 5
##       I     p sigma_ES sigma_fixed mean_nj
##   <dbl> <dbl>    <dbl>       <dbl>   <dbl>
## 1   0.1   0.5      0.2           1    11.1
## 2   0.2   0.5      0.2           1    25
## 3   0.3   0.5      0.2           1    42.9
## 4   0.4   0.5      0.2           1    66.7
## 5   0.5   0.5      0.2           1   100
## 6   0.6   0.5      0.2           1   150
## 7   0.7   0.5      0.2           1   233.
## 8   0.8   0.5      0.2           1   400
## 9   0.9   0.5      0.2           1   900
```

The plot below shows the relationship between I and the required site size for different levels of the cross-site impact variation in effect size units (`sigma_ES`) and the proportion of unites treated (`p`).

```
# Create a graph depicting the results
ggplot(data = df_result, aes(x = I, y = mean_nj,
                             color = factor(sigma_ES),
                             shape = factor(sigma_ES))) +
  geom_point() + geom_line() +
  geom_text(aes(label = round(mean_nj, 0)),
            size = 3.5, hjust = 0.0, vjust = -0.5) +
  facet_wrap(~p, labeller = "label_both") +
  scale_x_continuous(limits = c(0, 1.0),
                     breaks = seq(from = 0.0, to = 1.0, by = 0.1)) +
  scale_y_continuous(limits = c(0, 2000),
                     breaks = seq(from = 0, to = 2000, by = 200)) +
  scale_color_manual(values = c("dodgerblue1", "firebrick1")) +
  theme_bw() +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.direction = "horizontal") +
  labs(title = "Implied mean site size by a given level of information (I)",
       subtitle = "by cross-site impact SD in effect size units & proportion of units treated",
       y = "Implied mean site size",
```

```
       x = "Average reliability of the ML estimates (I)")
```



Major visible trends are as follows:

1. Implied mean site size **exponentially** goes up with the higher level of information (I).
2. Implied mean site sizes are higher for lower cross-site impact SD in effect size units (`sigma_ES`).
3. When the proportion assigned to treatment (`p`) deviates away from 0.50, implied mean site size goes up.

Does this plot show that our previous choice of the average reliability, $I = \{0.1, 0.5, 0.9\}$, match poorly with the actual settings of multisite trials in education research?

Assuming `sigma_ES = 0.20` and `p = 0.50`, for example, we have $\text{GM}(n_j) = 11.1$ for $I = 0.1$, which seems quite small. If we set $I = 0.9$, we obtain $\text{GM}(n_j) = 900$. This number seems quite large relative to actual site sizes we encounter from real-world multisite trials.

6 Actual site sizes from Weiss et al. (2017)

From Table 1 in Weiss et al. (2017), we can calculate a simple average site size for each multisite trial by dividing the total sample size (N) by the total number of sites (J).

```
# Calculate a simple average site size from Table 1 of Weiss et al. (2017)
weiss_df1 <- weiss_df1 %>%
  mutate(mean_site_size = N_persons/J_sites,
         mean_block_size = N_persons/RA_blocks)

# Display summary statistics
head(weiss_df1)
```

```
##   N_persons J_sites RA_blocks prop_treat mean_site_size mean_block_size
```

```
## 1      3566     316       316       0.63       11.28481        11.28481
## 2      3593     317       317       0.62       11.33438        11.33438
## 3      3531     317       317       0.63       11.13880        11.13880
## 4      3586     317       317       0.62       11.31230        11.31230
## 5      3601     318       318       0.61       11.32390        11.32390
## 6      3616     318       318       0.62       11.37107        11.37107
```

```
summary(weiss_df1$J_sites)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   17.00   34.00   77.63  100.00  318.00
```

```
summary(weiss_df1$RA_blocks)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    20.0    36.0    78.0   111.9   114.5   356.0
```

```
summary(weiss_df1$mean_site_size)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.14   75.06  102.38  188.82  163.33 1176.25
```

```
summary(weiss_df1$mean_block_size)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.14   41.56   70.33   99.34  102.18 1176.25
```

This simple approximate site sizes range from 11.3 (Head Start Impact Study, HSIS) to 1,176 (Welfare-to-Work Program, WtW), with most values around 100.

Here is a specific example from HSIS, further considering `sigma_ES` and `p`:

```
# (1) Head Start Impact Study (HSIS)
# N = 3,593, J = 317, p = 0.62, sigma_ES = 0.30 (for early reading score)
# simple average site size = 11.3
params <- list(
  I = seq(from = 0.1, to = 0.9, by = 0.1),
  p = 0.62,
  sigma_ES = 0.30,
  sigma_fixed = 1.0
) %>% cross_df()

GM_nj <- pmap(params, GM_site_size)

params %>%
  mutate(mean_nj = GM_nj) %>%
  unnest(cols = mean_nj) %>%
  mutate(mean_nj = round(mean_nj, 1)) %>%
  select(-sigma_fixed)
```

```
## # A tibble: 9 x 4
##       I     p sigma_ES mean_nj
##   <dbl> <dbl>    <dbl>   <dbl>
## 1   0.1  0.62      0.3     5.2
## 2   0.2  0.62      0.3    11.8
## 3   0.3  0.62      0.3    20.2
## 4   0.4  0.62      0.3    31.4
## 5   0.5  0.62      0.3    47.2
## 6   0.6  0.62      0.3    70.7
```

```
## 7   0.7  0.62      0.3   110
## 8   0.8  0.62      0.3   189.
## 9   0.9  0.62      0.3   424.
```

Note that we expect approximately $I = 0.2$ in the HSIS context, with the given value of simple average site size (11.3). In other words, the implied level of within-site information that can be achieved with the rough average site size of 11.3 in HSIS is about 0.2.

Here is one more example from Learning Coummunities (LC) study:

```
# (2) Learning Communities (LC) study
# N = 6,974, J = 11, p = 0.57, sigma_ES = 0.20 (for cumulative target credits earned)
# simple average site size = 634
params <- list(
  I = seq(from = 0.1, to = 0.9, by = 0.1),
  p = 0.57,
  sigma_ES = 0.20,
  sigma_fixed = 1.0
) %>% cross_df()

GM_nj <- pmap(params, GM_site_size)

params %>%
  mutate(mean_nj = GM_nj) %>%
  unnest(cols = mean_nj) %>%
  mutate(mean_nj = round(mean_nj, 1)) %>%
  select(-sigma_fixed)
```

```
## # A tibble: 9 x 4
##       I     p sigma_ES mean_nj
##   <dbl> <dbl>    <dbl>   <dbl>
## 1   0.1 0.570      0.2    11.3
## 2   0.2 0.570      0.2    25.5
## 3   0.3 0.570      0.2    43.7
## 4   0.4 0.570      0.2    68
## 5   0.5 0.570      0.2   102
## 6   0.6 0.570      0.2   153
## 7   0.7 0.570      0.2   238
## 8   0.8 0.570      0.2   408
## 9   0.9 0.570      0.2   918
```

Even with the quite large average site size (636), LC study cannot achieve $I = 0.9$ given `sigma_ES = 0.20` and `p = 0.5`.

Hence, $I = \{0.2, 0.5, 0.8\}$ seems more reasonable choice than $I = \{0.1, 0.5, 0.9\}$ for the simulation setup emulating real-world multisite trials. $I = 0.1$ and $I = 0.9$ look a bit extreme, considering the implied/required meann site sizes.

This is more evident when we simulate a vector of actual site sizes considering their heterogeneity across sites.

```
# Define a function to simulate site sizes
gen_site_sizes <- function(J = 25, I = 0.5, R = 1,
                           p = 0.50, sigma_ES = 0.20, sigma_fixed = 1.0){

  # Var(Y) = the ratio between 1) fixed sigma and 2) sigma in effect size units
  varY <- sigma_fixed^2/sigma_ES^2

  # Calculate the average level of site sizes implied by I
```

```
  kappa <- varY/(p*(1 - p))
  GM_se2_inv <- I/(sigma_fixed^2 * (1 - I))
  GM_nj <- kappa*GM_se2_inv

  # Generate nj_max and nj_min
  nj_max <- R*GM_nj
  nj_min <- (1/R)*GM_nj

  # Generate a site size vector
  nj_vec <- exp(seq(from = log(nj_min), to = log(nj_max), length = J))
  # N_vec <- seq(from = N_min, to = N_max, length = J)
  return(nj_vec)
}

# Get a site size vector for varying I levels
# given R = 2, J = 25
# "R = 2" means that the largest site size is two times the smallest site size
vec_I <- c(0.1, 0.2, 0.5, 0.8, 0.9)

params <- list(
  J = 25,
  I = vec_I,
  R = 2,
  p = 0.50,
  sigma_ES = 0.20,
  sigma_fixed = 1.0
) %>% cross_df()

(vec_nj <- pmap(params, gen_site_sizes) %>%
  set_names(paste0("I = ", vec_I)) %>%
  map(round, 0))
```

```
## $`I = 0.1`
##  [1]  6  6  6  7  7  7  8  8  9  9 10 10 11 12 12 13 14 15 16 17 18 19 20 21 22
##
## $`I = 0.2`
##  [1] 12 13 14 15 16 17 18 19 20 21 22 24 25 26 28 30 31 33 35 37 40 42 45 47 50
##
## $`I = 0.5`
##  [1]  50  53  56  59  63  67  71  75  79  84  89  94 100 106 112 119 126 133 141
## [20] 150 159 168 178 189 200
##
## $`I = 0.8`
##  [1] 200 212 224 238 252 267 283 300 317 336 356 378 400 424 449 476 504 534 566
## [20] 599 635 673 713 755 800
##
## $`I = 0.9`
##  [1]  450  477  505  535  567  601  636  674  714  757  802  849  900  954 1010
## [16] 1070 1134 1201 1273 1348 1429 1514 1604 1699 1800
```

7 The heterogeneity of the first-stage sampling errors (SEs)

In the original manuscript, the simulation factor R was defined as the ratio of the largest to smallest $\widehat{\text{se}}_j^2$ as in Paddock et al. (2006).

Luke pointed out the issue around how R is defined, noting that "in effect, as R goes up, I think our implied total sample size is also going up. i.e., the reduction in MSEL is driven by the implied overall increase in precision in our sites, rather than the heterogeneity."

Let's check this out with our new site size simulated provided above.

```
# Get a site size vector for varying R levels R = {1, 2, 5, 10}
# given I = 0.5, J = 25
# "R = 2" means that the largest site size is two times the smallest site size
vec_R <- c(1, 2, 5, 10)

params <- list(
  J = 25,
  I = 0.5,
  R = vec_R,
  p = 0.50,
  sigma_ES = 0.20,
  sigma_fixed = 1.0
) %>% cross_df()

vec_nj <- pmap(params, gen_site_sizes) %>%
  set_names(paste0("R = ", vec_R))

vec_nj %>%
  map(round, 0)
```

```
## $`R = 1`
##  [1] 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100
## [20] 100 100 100 100 100 100
##
## $`R = 2`
##  [1]  50  53  56  59  63  67  71  75  79  84  89  94 100 106 112 119 126 133 141
## [20] 150 159 168 178 189 200
##
## $`R = 5`
##  [1]  20  23  26  30  34  39  45  51  58  67  76  87 100 114 131 150 171 196 224
## [20] 256 292 334 382 437 500
##
## $`R = 10`
##  [1]   10   12   15   18   22   26   32   38   46   56   68   83  100  121  147
## [16]  178  215  261  316  383  464  562  681  825 1000
```

Actual vectors look like this.

```
# Arithmetic (additive) means
vec_nj %>%
  map(mean)
```

```
## $`R = 1`
## [1] 100
##
## $`R = 2`
```

```
## [1] 108.9029
##
## $`R = 5`
## [1] 153.7701
##
## $`R = 10`
## [1] 227.2096
```

```
# Geometric (multiplicative) means
geom_mean <- function(x){exp(mean(log(x)))}

vec_nj %>%
  map(geom_mean)
```

```
## $`R = 1`
## [1] 100
##
## $`R = 2`
## [1] 100
##
## $`R = 5`
## [1] 100
##
## $`R = 10`
## [1] 100
```

The error was an artifact of using multiplicative means instead of additive means.

What do we want R to index, in terms of our simulation?

$$\text{GM}(\widehat{\text{se}}_j^2) = \frac{\text{Var}(Y)}{p \cdot (1-p)} \cdot \text{GM}\left(\frac{1}{n_j}\right) = \kappa \cdot \text{GM}(n_j)^{-1}$$

```
# Define a function to generate a vector of simulated sampling errors
# Based on the simulated site sizes
gen_SE_vec <- function(J = 25, I = 0.5, R = 1,
                       p = 0.50, sigma_ES = 0.20, sigma_fixed = 1.0){

  # (1) ----- Simulate a vector of (effective) site sizes -----

  # Var(Y) = the ratio between 1) fixed sigma and 2) sigma in effect size units
  varY <- sigma_fixed^2/sigma_ES^2

  # Calculate the average level of site sizes implied by I
  kappa <- varY/(p*(1 - p))
  GM_se2_inv <- I/(sigma_fixed^2 * (1 - I))
  GM_nj <- kappa*GM_se2_inv

  # Generate nj_max and nj_min
  nj_max <- R*GM_nj
  nj_min <- (1/R)*GM_nj

  # Generate a site size vector
  nj_vec <- exp(seq(from = log(nj_min), to = log(nj_max), length = J))

  # (2) ----- Calculate SEs from the simulated site size vector -----
```

```
  SE_vec <- kappa*(1/nj_vec)  # scaled back by kappa
  return(SE_vec)
}
```

```
gen_SE_vec(J = 25, I = 0.1, R = 1, p = 0.20, sigma_ES = 0.10, sigma_fixed = 1.0)
```

```
##  [1] 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9
```

References

Miratrix, Luke W, Michael J Weiss, and Brit Henderson. 2021. "An Applied Researcher's Guide to Estimating Effects from Multisite Individually Randomized Trials: Estimands, Estimators, and Estimates." *Journal of Research on Educational Effectiveness* 14 (1): 270–308.

Rosenbaum, Paul R, and Donald B Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Weiss, Michael J, Howard S Bloom, Natalya Verbitsky-Savitz, Himani Gupta, Alma E Vigil, and Daniel N Cullinan. 2017. "How Much Do the Effects of Education and Training Programs Vary Across Sites? Evidence from Past Multisite Randomized Trials." *Journal of Research on Educational Effectiveness* 10 (4): 843–76.