AUGUST 2022

NIH Diversity Program Consortium

Technical Report

Enhance Diversity Study: Power Calculations for Student Outcome Analyses

Prepared by: Catherine M. Crespi of the Coordination and Evaluation Center of the Diversity Program Consortium, UCLA

www.diversityprogramconsortium.org

DIVERSITY PROGRAM CONSORTIUM
Supported by the National Institutes of Health

Coordination & Evaluation Center | UCLA

Introduction

The Building Infrastructure Leading to Diversity (BUILD) initiative of the National Institutes of Health (NIH) Diversity Program Consortium, which is funded by the NIH Common Fund and managed by the National Institute of General Medical Sciences (NIGMS), was established to incentivize undergraduate institutions to create innovative approaches to increasing diversity in biomedical research, with the ultimate goal of diversifying the NIH-funded research enterprise. An evaluation of the BUILD initiative is being implemented by the Coordination and Evaluation Center (CEC).

A major component of the CEC evaluation is the Enhance Diversity Study, a large-scale, systemic, national longitudinal evaluation of BUILD training programs (McCreath et al., 2017). The Enhance Diversity Study includes the systematic collection of qualitative and quantitative data from students and faculty to measure psychosocial factors and outcomes. Consortium-wide data are collected at defined intervals, including participant rosters for BUILD activities, student and faculty survey responses, institutional records, and transcripts from CEC case studies.

Adequate power for statistical analyses to estimate the impact of BUILD involvement on student outcomes is crucial for successfully generating recommendations for policy and practice. In this technical report, we report results of power calculations conducted to determine the smallest differences in outcomes between BUILD-exposed students and students not involved in BUILD that can be detected with 80% power, given the sample sizes of students at BUILD institutions that are expected to be available for longitudinal analysis. Power of 80% is a commonly used benchmark and represents a probability of 0.80 of concluding that an intervention effect exists, given that such an effect exists.

METHODS

Definition of BUILD Students

For purposes of this technical report, “BUILD” student is defined as being a BUILD Scholar, a BUILD Associate, and/or having a BUILD program undergraduate research experience (URE). BUILD Scholar is program-defined. These are the most intensely treated and supported group of students. Scholars often receive tuition support or stipend, research training, and mentorship. Compulsory and structured participation in a host of BUILD activities is common for this group. BUILD Associate is also program defined. This term describes a less intensely treated and supported group of students, often participating in a subset of structured BUILD activities. Some programs recruit Scholars from the Associate pool. URE is reported by programs and includes BUILD affiliated student-directed research and/or mentored undergraduate research experiences, during the academic and/or summer term(s).

Power Calculation Approach

Our objective was to answer the question, given the sample sizes that are expected to be available for longitudinal analysis, what are the smallest effect sizes that can be detected with 80% power and a two-sided significance level of 0.05? Since most outcomes of interest for BUILD students are either continuous variables (e.g., scale variables such as Science Identity) or dichotomous variables (e.g., persistence in a biomedical major), we sought to calculate the smallest detectable effect sizes for a continuous outcome and a dichotomous outcome.

Effect Sizes

For continuous outcomes, the effect size was operationalized as the standardized mean difference between BUILD and non-BUILD students at follow-up. The standardized mean difference is a widely used effect size measure for continuous outcomes and is equal to

$$\frac{Mean_{BUILD} - Mean_{non\text{-}BUILD}}{SD}$$

where SD is the standard deviation of the outcome variable. The standardized mean difference is in units of standard deviation. Commonly accepted benchmarks are that standardized mean differences of 0.2 and 0.5 represent small and medium effect sizes (Cohen 1988).

For the dichotomous outcomes, the effect size is the difference in proportions,

$$P_{BUILD} - P_{non\text{-}BUILD}$$

Benchmarks for differences in proportions that represent small, medium and large effect sizes have been proposed (Cohen, 1988). In general, differences in proportions on the order of 0.05-0.10 represent small effect sizes and differences on the order of 0.15-0.25 represent medium effect sizes.

Factor Influencing Power

The following factors influence power and were part of the calculations:

(1) **Multilevel design:** Data collected to evaluate the BUILD initiative are multilevel, with sites at the upper level and students within sites at the lower level. Our calculations were based on power methods for studies with multilevel designs (Moerbeek & Teerenstra, 2016). In a multilevel design with two levels such as BUILD, variation in an outcome variable arises due to variation at two levels: variation across sites and variation of participants within sites. Power calculations require plausible estimates of the variances at the two levels or the intraclass correlation coefficients that quantify the apportionment of variance between levels. Where possible, we obtained plausible estimates of these quantities by fitting models to data. This modeling is explained further below.

(2) Sample size: number of sites (upper-level units): There are 10 BUILD programs, with 11 primary institutional sites for data collection. Different numbers of sites are expected to be available for inclusion in different analyses. Of note, institutional records data will be available for a limited number of students for two sites due to consent requirements at these institutions. At these two sites, only students who granted permission to the institution on a survey in 2020, 2021, or 2022 will contribute to analyses. Thus, we performed computations under two scenarios: 11 sites and 9 sites.

(3) Sample size: number of students per site (lower-level units): To obtain plausible estimates of the number of BUILD students per site that would be included in analyses, we determined that approximately 1000 BUILD students have completed a baseline survey and at least two follow up surveys and approximately 1500 BUILD students have completed a baseline survey and at least one follow-up survey. These numbers correspond to means of 91 and 136 BUILD students per site, respectively. Numbers of students with outcomes based on institutional records may be somewhat lower. We present results for a mean number of BUILD students per site ranging from 40 to 120, since sample sizes may vary from analysis to analysis due to differences in survey response rates, availability of institutional records, missingness, and focus on subgroups

(4) Heterogeneity of the treatment effect: In a multilevel design in which there are both intervention-exposed individuals and comparison individuals at each site, the intervention effect may vary from site to site. This is referred to as heterogeneity of the treatment effect and tends to reduce power (Moerbeek and Teerenstra, 2016). Our calculations allow for plausible levels of heterogeneity of the treatment effect. Where possible, we obtained estimates of the between-site variance of the treatment effect or the corresponding intraclass correlation coefficient by fitting statistical models to BUILD study data. This modeling is explained further below.

(5) Variability in sample size per site: In the BUILD initiative, the number of students varies from site to site. Variation in the number of observations per site in a multilevel study tends to reduce power. We discounted the sample sizes by 5% to account for this loss of power, based on literature indicating that the impact of varying cluster/site sizes rarely exceeds a 10% efficiency loss (van Breukelen et al, 2007).

(6) Ratio of BUILD to non-BUILD students at each site: At most BUILD sites, the number of non-BUILD students who could be considered a comparator group for BUILD students (i.e., biomedical science majors), is much larger than the number of BUILD students. However, available power methods for multisite studies assume equal numbers of intervention and control participants per site. Therefore our calculations are restricted to assuming equal numbers of BUILD and non-BUILD students per site. This assumption is conservative in that it entails an underestimation of the actual sample size that is available.

(7) Impact of confounding: Standard power calculation methods are for randomized trials. Since the BUILD initiative is not a randomized trial, calculations assuming randomization will tend to be overestimate the power of the study. Adjusting for patterns of confounding in sample size/power calculations for observational or quasi-experimental studies is extremely challenging, both technically and scientifically. Some studies have shown that naïve methods of sample size calculation can underestimate the required sample size by as much as half (Haneuse et al. 2012). As an adjustment for the impact of confounding, we discounted the sample sizes by 20%.

Variance Parameter Estimates

To conduct power calculations for a study with a multilevel design such as the BUILD initiative, it is necessary to have estimates of variance parameters. Where possible, we fit models to data to obtain plausible estimate of the necessary parameters. The necessary parameters and the statistical procedures to estimate them are somewhat different for continuous and dichotomous outcomes. We explain the procedures for each.

Continuous outcomes. For continuous outcomes, there were three relevant variance parameters: the variance of site-level means, the variance of the intervention effect across sites (relevant to heterogeneity of the treatment effect), and the residual variance at the individual level. To obtain estimates of these parameters, we fit models using two illustrative outcomes: science self-efficacy and science identity. These variables are collected on student surveys and are scored using item response theory. Linear mixed models were fit to data from students who completed a baseline survey and at least one follow-up survey and were biomedical science majors at the last survey. The dependent variable was the outcome at follow-up. Models included an indicator for BUILD exposure, controlled for the outcome variable at the baseline (first) survey, and included a random intercept for site and a random effect for the BUILD effect (by site).

After obtaining the variance parameter estimates from the models, we then used them to calculate intraclass correlation coefficients (ICCs). ICCs are quantities for multilevel data that characterize how the variance of the outcome is apportioned between the upper (site) level and the lower (student) level. It is the ICC estimates that are used in the power calculations. The two relevant ICCs are ICC_0, which is the proportion of the variance that is due to variation of mean outcomes across sites, and ICC_1, which is the proportion of the variance that is due to heterogeneity of the treatment effect across sites.

Estimates of the ICCs obtained from fitting models to data are presented in Table 1. ICC_0 quantifies the proportion of the total variance of an outcome that is attributable to variation of site-level means, and as it increases, power increases. ICC_1 quantifies the proportion of the total variance of an outcome that is attributable to variation of the treatment effect across sites, and as it increases, power is reduced. The impact of ICC_1 tends to be larger than that of ICC_0. Raudenbush and Liu (2000) have proposed 0.05, 0.10, and 0.15 as small, medium, and large values of ICC_1. Thus the ICC values in Table 1 can be considered very small. This implies that the multilevel nature of the BUILD data has a relatively small impact on power.

Table 1. Intraclass correlation coefficient estimates for continuous outcomes obtained by fitting models to longitudinal data from the BUILD initiative

	Science Self-Efficacy	Science Identity
ICC_0: proportion of the variance that is due to variation of mean outcomes across sites	0.0063	0.0079
ICC_1: proportion of the variance that is due to heterogeneity of the treatment effect across sites	0.0088	0.0099

For purposes of calculating the smallest detectable effect size, we assumed an ICC_0 of 0.006 (obtained by rounding the lower value down) and an ICC_1 of 0.01 (obtained by rounding the higher value up).

Dichotomous outcomes. For dichotomous outcomes, we will be comparing outcome proportions between BUILD and non-BUILD comparison students. A statistical feature of proportions is that the variance of an estimated proportion depends on the value of the proportion. Proportions near 0.5 have the highest variance and proportions close to 0 or 1 have the lowest variance. Thus, power is the lowest when trying to detect differences in proportions when the proportions are near 0.5 and highest when the proportions are near 0 or 1.

In order to conduct power calculations, it is necessary to specify the values of the proportions in each group, rather than just the difference in proportions. We selected an outcome proportion of 0.75 for the comparison group and assumed that the proportion would be higher in the BUILD group; the smallest difference in proportions that could be detected was the minimal detectable effect size. The choice of 0.75 was based on data showing that persistence in biomedical majors was about 0.75 among incoming biomedical majors not exposed to BUILD. This proportion also reflects a somewhat but not overly conservative choice.
For dichotomous outcomes for the evaluation, the data will be modeled using multilevel logistic regression models. In these models, the effect of BUILD is estimated as a log odds ratio, and the variance parameters are also on the log odds scale. There were two variance parameters: the variance of the log odds of the outcome across sites, and the variance of the intervention effect across sites (due to treatment effect heterogeneity). An estimate of the variance of the log odds of the outcome across sites can be obtained based on the specified outcome proportions in each group. While it is theoretically possible to obtain an estimate of the variance of the intervention effect across sites by fitting models to data, currently available data on persistence did not provide a reliable estimate of this parameter due to sparseness of data from some sites. Therefore we selected a plausible value for the variance of the intervention effect across sites using the approach suggested by Moerbeek and Teerenstra (2016, page 127). This approach involves selecting a value for the variance that will yield a plausible interval within which we might expect 95% of the site-level intervention effects to lie. This value was determined to be 0.05.

RESULTS

The smallest effect sizes that can be detected with 80% power, based on the assumptions and procedures described in the Methods section, are presented in Tables 2 and 3. Results are shown for a range of mean number of BUILD students per site. Table 2 is applicable to analyses that include all 11 primary institutional sites and Table 3 applies to analyses including only nine sites. Many analyses utilizing student survey responses are expected to include all 11 sites, whereas analyses that rely on the availability of institutional records may be limited to nine sites in some instances.

Table 2. Smallest effect sizes detectable with 80% power: analyses including 11 sites

Mean number of BUILD students per site	Continuous outcome: standardized mean difference	Dichotomous outcome: difference in proportions
120	0.27	0.086
100	0.28	0.09
80	0.30	0.10
60	0.33	0.11
40	0.39	0.14

Table 3. Smallest effect sizes detectable with 80% power: analyses including 9 sites

Mean number of BUILD students per site	Continuous outcome: standardized mean difference	Dichotomous outcome: difference in proportions
120	0.31	0.095
100	0.32	0.10
80	0.34	0.11
60	0.38	0.13
40	0.44	0.15

The results show that, when conducting longitudinal analyses that involve a mean of 100-120 BUILD students per site, we will have power of at least 80% to detect effect sizes in the small range. As mentioned in the Methods, about 1000 BUILD students, or about 91 per site, have completed a baseline survey and at least two follow up surveys, and about 1500 BUILD students, or about 136 per site, have completed a baseline survey and at least one follow-up survey. Thus, these estimates of detectable effect sizes for the case of 100-120 BUILD students per site can be considered applicable to many consortium-wide analyses.

The tables show that, as the mean number of BUILD students per site in the analysis decreases, the smallest effect sizes that can be detected with 80% power increase. In general, these effect sizes are in the small-to-medium range. Smaller numbers of BUILD students per site may be available for some analyses due to missing data or because the analyses focus on subgroups.

DISCUSSION

The goal of the analyses presented in this technical report was to determine the smallest differences in outcomes between BUILD-exposed and non-BUILD students that are detectable with 80% power, given the sample sizes that are expected to be available for longitudinal analysis. It is important to have adequate power to detect meaningful differences between groups, since inadequate power can lead to important effects going undetected and the erroneous conclusion that a program or intervention was not effective.

We found that longitudinal analyses using consortium-wide data should have adequate power to detect small effect sizes, on the order of a standardized mean difference of 0.3 for continuous outcomes and a difference of proportions on the order of 0.09-0.10. Such differences are meaningful and are consistent with effects that we have been detecting in analyses. Our calculations involving smaller numbers of BUILD students per site, e.g., 40-60 per site, may be considered relevant to subgroup analyses that are restricted to individuals with certain characteristics. For such analyses, we have adequate power to detect small-to-medium effect sizes, on the order of 0.35-0.40 for standardized mean differences and 0.11-0.15 for differences in proportions. However, the calculations here do not consider power for analyses that compare outcomes across different subgroups.

The calculations in this report have limitations. Conducting power calculations involves making assumptions about the values of various parameters. We attempted to make plausible assumptions about the values of parameters based on currently available information. In general, we made assumptions that were somewhat conservative. We were also limited by the power calculation techniques that are currently available. Our calculations were based on power methods designed for randomized trials with multilevel data that have equal numbers of "treated" and "untreated" participants at each site. Therefore, our calculations do not account for the fact that many analyses are likely to include larger numbers of non-BUILD students, rather than equal numbers of BUILD and non-BUILD students. The assumption of equal numbers is conservative in that it is an underestimation of the actual sample size. Overall, due to the mostly conservative assumptions, the minimum detectable effect sizes presented here should be considered high; it is likely that we will have adequate power to detect smaller effect sizes.

Challenges and efforts to sustain engagement with the nationwide longitudinal study have been well described. Analyses are underway to better understand sample representativeness and the potential bias of non-response. Nonresponse bias may occur when respondents to a survey differ from non-respondents. This issue is the focus of separate analyses comparing the EDS sample to the Integrated Postsecondary Education Data System (IPEDS) data for these institutions. Findings from these analyses will be made available in a forthcoming technical report.

Author Details

Catherine M. Crespi, Ph.D.
Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, CA 90095, USA.

Suggested Citation for the Technical Report:

Crespi, C.M. (2022). TECHNICAL REPORT: Enhance Diversity Study: Power Calculations for BUILD Student Outcome Analyses. Los Angeles, CA: Diversity Program Consortium Coordination and Evaluation Center at UCLA.

Acknowledgements

The author of this report would like to acknowledge the many students and faculty who generously shared their time and experience for the purposes of Diversity Program consortium (DPC). The author also wishes to acknowledge the contribution to this work from Claudia Perez, a colleague of the CEC, who produced the data set for power calculations.

References

Cohen J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Haneuse S, Schildcrout J, Gillen D. (2012). A two-stage strategy to accommodate general patterns of confounding in the design of observational studies. Biostatistics 13(2):274-288.

McCreath HE, Norris KC, Calderón NE, Purnell DL, Maccalla NM, Seeman TE. Evaluating efforts to diversify the biomedical workforce: the role and function of the Coordination and Evaluation Center of the Diversity Program Consortium. BMC Proceedings 11(12):15-26.

Moerbeek M, Teerenstra S. (2016). Power Analysis of Trials with Multilevel Data. CRC Press/ Taylor & Francis Group.

Raudenbush SW, Liu X. (2000). Statistical power and optimal design for multisite randomized trials. Psychological Methods 5(2): 199-213.

Van Breukelen GJ, Candel MJ, Berger MP. (2007). Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicenter trials. Statistics in Medicine 26(13):2589–2603.

The Diversity Program Consortium's Coordination and Evaluation Center at UCLA, and the Enhance Diversity Study, is supported by the Office of the Director of the National Institutes of Health/ National Institutes of General Medical Sciences under award numbers U54GM119024 and U54GM119024–03:S1.

www.diversityprogramconsortium.org

Diversity Program Consortium
Supported by the National Institutes of Health

Coordination & Evaluation Center | UCLA