Analytic Technical Assistance and Development
U.S. Department of Education
July 2017

What is design-based causal inference for RCTs and why should I use it?

Peter Z. Schochet
Mathematica Policy Research, Inc.

NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences
U.S Department of Education

NCEE 2017–4025

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased, large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

July 2017

This report was prepared for the Institute of Education Sciences (IES) by Decision Information Resources, Inc. under Contract ED-IES-12-C-0057, Analytic Technical Assistance and Development. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Contents

Boxes

(This page intentionally left blank)

Design-based causal inference—a recent approach for impact estimation

Design-based methods have recently been developed as a way to analyze data from impact evaluations of interventions, programs, and policies (see, for example, Imbens and Rubin, 2015 and Schochet, 2015, 2016). The impact estimators are derived using the building blocks of experimental designs with minimal assumptions, and have good statistical properties. The methods apply to randomized controlled trials (RCTs) and quasi-experimental designs (QEDs) with treatment and control (comparison) groups. Importantly, design-based estimators are acceptable for What Works Clearinghouse (WWC) evidence reviews (Scher and Cole, 2017). The free *RCT-YES* software (www.rct-yes.com) estimates impacts using these estimators, and reports impact findings in formatted tables with key information required by the WWC.[1]

Although the fundamental concepts that underlie design-based methods are straightforward, the literature on these methods is technical, with detailed mathematical proofs required to formalize the theory. Thus, the daunting task of wading through this literature might discourage some education researchers from using design-based methods in favor of more traditional "model-based" methods, such as hierarchical linear modeling (HLM) (Raudenbush and Bryk, 2002).

This brief aims to broaden knowledge of design-based methods by providing intuition on their key concepts and how they compare to model-based methods as typically implemented. Using simple mathematical notation, the brief is geared toward education researchers with a good knowledge of evaluation designs and HLM. The discussion synthesizes Schochet (2016), omitting details for brevity and accessibility. The focus is on RCTs, although key concepts apply also to QEDs.

The theory in a nutshell

Design-based theory is derived directly from the Neyman-Holland-Rubin causal inference model that underlies experiments (Holland, 1986; Rubin, 1974). Consider the simplest RCT design where individuals are randomly assigned to either a treatment group that is offered an intervention or a control group that is not. The study participants are followed for a period of time and outcome data, such as achievement test scores, are collected on the sample. Let y_i denote the outcome variable for individual *i*.

Ideally, we would like to measure each individual's "potential" outcome in the treatment condition (Y_{Ti}) and in the control condition (Y_{Ci}). With this information, we could calculate each individual's treatment effect, $(Y_{Ti} - Y_{Ci})$, and then the average treatment effect, $\beta_{ATE} = \overline{Y}_T - \overline{Y}_C$, which is the impact parameter of interest for most evaluations in the education field.

[1] *RCT-YES* was funded by the Institute of Education Sciences (IES) to support the conduct of rigorous impact evaluations by education agencies and school districts.

Unfortunately, we can only observe either Y_{Ti} or Y_{Ci} depending on the random assignment results, but not both. This means we cannot directly calculate individual and average treatment effects. We can demonstrate this concept mathematically by relating the observed outcome, y_i, to the potential outcomes, Y_{Ti} and Y_{Ci}, as follows:

$$(1) \quad y_i = T_i Y_{Ti} + (1 - T_i) Y_{Ci},$$

where T_i is an indicator variable that equals 1 for those assigned to the treatment group and 0 for those assigned to the control group. Equation (1) simply states we can observe $y_i = Y_{Ti}$ for those in the treatment group and $y_i = Y_{Ci}$ for those in the control group.

Design-based theory uses the relation in Equation (1) to develop estimators for the unobserved average treatment effect, β_{ATE}. The idea is to first add $T_i \overline{Y}_T$ and $(1 - T_i)\overline{Y}_C$ to both sides of Equation (1)—which does not change the equation—and to then rearrange terms in the equation to produce the following regression model:

Equations (1) and (2) underlie design-based theory by linking the observed data to the random assignment process. All impact estimators are derived using these relations, with minimal assumptions.

$$(2) \quad y_i = \beta_0 + \beta_{ATE} T_i + u_i,$$

where $\beta_0 = \overline{Y}_C$ is the intercept, $\beta_{ATE} = \overline{Y}_T - \overline{Y}_C$ is the average treatment effect parameter we want to estimate, and u_i is the model "error" term. Importantly, u_i is random only because T_i is random; the potential outcomes are assumed to be fixed for the study.

The design-based model in Equation (2) has statistical properties that differ from the standard linear model typically used to estimate impacts for RCTs. For example, the error term, u_i, does not have mean 0 or constant variance and is correlated with the regressor, T_i. Yet it can be shown that estimating this model using standard ordinary least squares (OLS) produces a differences-in-means impact estimator based on the observed data, $\hat{\beta}_{ATE} = \overline{y}_T - \overline{y}_C$, that has the following desirable statistical properties (see, for example, Schochet, 2016):

- ***Unbiased***, meaning that the estimator will, on average, equal the true impact parameter across all possible random assignment results
- ***Normally distributed in large samples***, so that standard *t*-tests or *z*-tests can be used to test the null hypothesis of zero average treatment effects
- ***Simple variance estimator***, shown in Box 1 at the end of the brief, with separate variance terms for the treatment and control groups

This same theory applies to impact estimators for the full sample and for subgroups defined by baseline characteristics, such as gender or academic proficiency level in the prior year.

The key feature of design-based theory, then, is that it uses the random assignment process to *build* the impact estimation model in Equation (2). In contrast, model-based approaches specify an ad hoc model structure (for example, the standard OLS model) that is assumed to be true to ensure unbiased estimators. But it is not possible to fully verify these model assumptions. We discuss differences between the design-based and model-based approaches more fully later in the brief.

Extensions to clustered (multilevel) designs

The above theory can be extended to clustered designs where groups (such as schools or classrooms) are randomly assigned to the treatment and control groups instead of individuals. For example, a common design used in education RCTs randomizes schools and collects outcome data for students.

For clustered designs, for simplicity, we consider *design-based methods that average the individual data to the group level*, although individual data could also be used for estimation (Schochet, 2013). As an example, in an RCT where schools are randomized, we consider estimators where the student-level data (such as test scores) are averaged to the school level for the analysis. In this context, we can define potential outcomes (student averages) for school j as $\bar{Y}_{Tj}$ for treatment schools and $\bar{Y}_{Cj}$ for control schools. The school-level treatment effect is $(\bar{Y}_{Tj} - \bar{Y}_{Cj})$, which cannot be observed because we can only measure $\bar{Y}_{Tj}$ or $\bar{Y}_{Cj}$, but not both. The impact parameter of interest, $\beta_{ATE,Clus} = \bar{\bar{Y}}_{TW} - \bar{\bar{Y}}_{CW}$, is a weighted average of these school-level treatment effects, which can also be expressed as a weighted average of student-level treatment effects (weights are discussed in more detail later).

Parallel to Equation (1) for the non-clustered design above, we can now relate the observed mean outcome, $\bar{y}_j$, to the potential outcomes, $\bar{Y}_{Tj}$ and $\bar{Y}_{Cj}$, as follows:

$$(3) \quad \bar{y}_j = T_j \bar{Y}_{Tj} + (1 - T_j)\bar{Y}_{Cj},$$

where T_j equals 1 for schools assigned to the treatment group and 0 for schools assigned to the control group. As before, we can add $T_j \bar{Y}_{Tj}$ and $(1 - T_j)\bar{Y}_{Cj}$ to both sides of this equation and rearrange terms to form a regression model similar to Equation (2) where $\bar{y}_j$ is regressed on T_j, with the model error term, u_j, defined by the randomization process:

Equations (3) and (4) underlie design-based theory for clustered designs where groups, such as schools, are randomized. These relations link the observed data—averaged to the cluster level—to the random assignment process.

$$(4) \quad \bar{y}_j = \beta_{0,Clus} + \beta_{ATE,Clus} T_j + u_j .$$

Estimating this model using weighted least squares yields a weighted differences-in-means impact estimator that, in large samples, is unbiased (consistent) and normally distributed with a simple variance estimator (see Box 2 at the end of the brief). Standard *z*-tests or *t*-tests can be used for hypothesis testing where the degrees of freedom are based on the number of clusters in the sample.

This approach to clustering differs from HLM, where an ad hoc model specification and error structure are assumed to be true at each HLM level. But it is difficult to fully verify these HLM model assumptions. Next, we discuss the advantages of the design-based approach in more detail.

Advantages of the design-based approach

Education researchers have typically used OLS or HLM methods to analyze RCT data. The main advantage these methods have over the design-based methods is that they could yield more precise impact estimates. However, this will only necessarily happen if the *model is specified correctly*. With misspecification, the model-based approaches could yield biased estimates. For example, HLM models for multilevel designs typically assume the error terms are additive at each level, normally distributed, independent of each other, and uncorrelated with the treatment status indicator variable. These models also typically assume treatment effects are the same for everybody. These model assumptions may or may not be correct; the design-based approach avoids our having to make them.

Benefits of the design-based approach include:

- Requires minimal assumptions
- Applies to continuous and binary outcomes
- Yields simple variance estimators
- Requires less data for clustered designs
- Allows assumptions on the generalizability of results
- Accommodates flexible models with baseline covariates
- Extends to blocked designs
- Provides transparency on cluster and block weighting
- More suited to RCTs than "robust" estimators
- Performs well in simulations

There are several important advantages to the design-based approach:

Requires minimal assumptions. The design-based approach relies only on the randomization mechanism to develop estimators, and thus relies on fewer assumptions. Design-based methods do not require assumptions on the distributions of potential outcomes, and thus are non-parametric. In addition, unlike typical model-based assumptions, treatment effects are allowed to vary from one individual to the next. The approach emphasizes "robust" inference that could be less sensitive to model misspecification.

There are three main assumptions required for the design-based estimators that are *also* required for the model-based estimators. The first is that potential outcome distributions have finite means and variances, which is likely to hold for most education outcomes. The second is that the potential outcomes of an individual depend only on that individual's treatment or control assignment and

not on the assignments of other individuals in the sample.[2] In the education context, this assumption is often plausible, but not always—for example, for designs where the treatment status of one student could affect the outcomes of other students in the same school or neighborhood due to peer effects. The final assumption is the independence between treatment status and potential outcomes, which is ensured by randomization for RCTs and is assumed to hold conditional on baseline covariates for QEDs with comparison groups.

Applies to different types of outcome variables. Design-based estimators apply to both continuous outcomes (such as achievement test scores) and binary (1/0) outcomes (such as whether a student dropped out of school). Thus, there is no need to estimate impacts for binary outcomes using more complex logit or probit models that are often used for model-based analyses.

Yields simple variance estimators. Estimators under the design-based approach are more transparent and easier to apply. The design-based approach yields explicit formulas for the impact and variance estimators, even for complex designs. In contrast, HLM methods require iterative, numerical maximum likelihood procedures that must converge.

Requires less data for clustered designs. For multilevel designs, the design-based estimators require data only on cluster averages (see Equation (4)), whereas HLM methods by default use data at the individual level. This difference has practical importance because education researchers are increasingly using administrative records as a primary data source for their impact evaluations. Thus, design-based methods can help researchers gain access to these records by allaying common concerns that data agencies have about the potential for unwanted data disclosure if they release individual records. For example, for a clustered RCT with school-level random assignment, study researchers would need only to request school-level averages for the full sample of students and for each subgroup of interest (for instance, separate school-level averages for girls and boys).

Allows for different assumptions about whether the impact findings can generalize beyond the study sample. The design-based approach allows the analyst to decide whether it is more realistic to assume (i) the "*finite-population model*" where the study results pertain only to study participants, sites, and intervention offerings at the time of the study or (ii) the "*super-population model*" where the study results generalize to a broader population. This choice might depend on the study context—for example, the number and range of sites, the purposeful or random selection of sites and individuals, and how the impact findings will be used for policy. The HLM approach, however, does not allow this choice: it assumes a super-population model.

The design-based theory presented thus far is a *finite-population* model where potential outcomes are assumed to be fixed for the study. Under the *super-population* framework, the potential outcomes in the regression models are instead assumed to be randomly sampled from a broader population

[2] The literature refers to this condition as the stable unit treatment value assumption (SUTVA) (Rubin, 1986).

(which may be vaguely defined). This leads to estimators with statistical properties similar to those for the finite-population model, except that the final term of the variance estimators in Boxes 1 and 2 is divided by the size of the super-population rather than the sample size. Thus, variances are *larger* under the super-population model, reflecting the statistical "penalty" of generalizing the impact findings to a broader population.

Accommodates models with baseline covariates without assuming they enter the model additively. Baseline covariates (such as pre-intervention test scores) are often included in impact estimation models to improve the precision of the impact estimates and to adjust for random imbalances between the treatment and control groups. Model-based methods typically assume these covariates enter the model additively. But how do we know this assumed relationship between the outcomes and covariates is valid?

Design-based methods do not require this assumption because the covariates do not enter the "true" RCT data-generating process in Equations (1) and (3). However, covariates can be added to the model in the typical way using a design-based variant of the OLS multiple regression estimator. This approach yields impact estimators that are consistent and asymptotically normal, with variance estimators similar to those in Boxes 1 and 2 except mean squared residuals from the fitted regression models are used in place of the s_T^2, s_C^2, s_{TW}^2, and s_{CW}^2 terms. Thus, this approach provides a principled framework for entering covariates into the model in the usual way without having to make assumptions about the relationship between the outcomes and covariates.[3]

Extends to blocked designs. Blocked designs are RCTs where random assignment is conducted *separately* within different sub-populations of the sample (for example, by site or grade level). The design-based approach uses this simple random assignment process in each block to develop impact estimators rather than specifying an ad hoc estimation model to account for the blocks. For the finite-population model, where the study blocks are treated as fixed for the study, the design-based estimators from above apply to each block separately, and can then be averaged to obtain overall impact findings. For the super-population model, where the study blocks are treated as a random sample, the form of the variance estimator differs but is still simple to apply in practice (see Box 3).

Provides transparency on how clusters and blocks are weighted for the analysis. An important but often overlooked analysis issue is the choice of weights for combining data across clusters (such as schools) and blocks (such as sites). This choice could partly depend on the study research questions, but could also depend on researcher concerns about the undue influence of very large sites on the overall impact findings. For example, in a clustered design with school-level randomization, schools could be weighted (i) equally, to obtain impacts for the average school in the sample; (ii) based on

[3] For clustered designs, the design-based models discussed in this brief can include group-level covariates but not individual-level ones, which can lead to some precision losses. However, this problem can be overcome using alternative design-based methods that use the individual-level data to estimate the regression models (Schochet, 2013).

student sample sizes, to obtain impacts for the average student in the sample; or (iii) using "precision" weighting, where schools whose mean student outcomes are measured more precisely are given larger weight in the analysis. By default, the HLM approach uses precision weighting.

An advantage of the design-based approach is its transparency with regard to how weights enter the analysis (see Boxes 2 and 3). This transparency can encourage researchers to select weights that best align with their key study questions and to conduct exploratory analyses to assess the sensitivity of the impacts to alternative weights. It is more difficult, however, to discern the weighting scheme for HLM estimators (especially for models that include baseline covariates) and to undo the default precision weighting scheme to implement alternative schemes.

Yields estimators that are similar in spirit to "robust" estimators, but have advantages for RCTs. It is common in certain fields to obtain standard errors for RCTs from OLS models that are robust to model misspecification. These estimators include robust standard errors for non-clustered designs (Huber, 1967; White, 1980) and extensions to clustered designs using generalized estimating equations (Liang and Zeger, 1986). These estimators share features with the design-based estimators. The benefit of the design-based approach, however, is that the randomization mechanism defines the model error terms. Thus, the variance estimators are derived directly from this *known* error structure. In contrast, the robust estimators provide variances for error structures that are *unknown*, and thus are not as tailored to the RCT context.

Performs well in simulations. The design-based estimators have similar statistical properties to the HLM and robust estimators in very large samples, but not necessarily in typical samples used in practice. Thus, to compare the statistical properties of the estimators in real-world applications, Schochet (2016) conducted simulations by (i) randomly generating many datasets from models with known impacts and variances and (ii) comparing the estimated impacts and variances across the simulations to their true (known) values. The simulations were conducted assuming a typical clustered education RCT with school-level randomization and a student test score outcome. The simulations were performed for designs with and without site-level blocking, for models with and without a pretest baseline covariate, and for various model specifications used to generate the data.

The simulation findings suggest that the design-based estimators are likely to perform well across a broad range of RCT designs used in education. For the clustered RCTs considered, estimated impacts have negligible bias if the sample contains at least 8 schools. Furthermore, with a sample of at least 12 schools, the standard errors produced by the design-based approach align with the true standard errors, and precision levels are comparable to those for the HLM and robust estimators, even for simulations based on the HLM assumptions.

Conclusions

Design-based methods provide a unified, principled framework for analyzing data from impact evaluations for a wide range of designs used in education research, and are a viable alternative to model-based approaches. The design-based approach uses the building blocks of experiments to derive impact estimators, and thus relies on fewer assumptions than model-based approaches. The design-based estimators are unbiased and normally distributed in large samples (facilitating hypothesis testing) and yield simple variance expressions even for complex designs and models with covariates (increasing transparency). The estimators for clustered designs require data only on cluster-level averages, which can help overcome concerns about data disclosure risks when collecting administrative records data. The approach also allows flexibility in the choice of estimators depending on whether the impact findings are assumed to pertain to the study sample only or to generalize more broadly. Finally, the design-based estimators perform well in simulations, suggesting they are likely to perform well in real-world impact evaluations.

Using data from nine education RCTs spanning a variety of different designs and contexts, Kautz, Schochet, and Tilley (2017) found that design-based, HLM, and robust estimators produced similar impact findings. Nonetheless, there were a few instances where impact estimates differed in statistical significance (that is, whether or not *p*-values were less than .05) across the three estimators. This occurred due to differences in underlying model assumptions, such as how blocks and clusters are weighted for the analysis and the choice of the finite- versus super-population model. Thus, it is critical for researchers to carefully consider their models and assumptions—regardless of the choice of estimator—to best identify appropriate analytic methods and statistical package options, rather than relying solely on default model specifications. Moreover, researchers should consider the tradeoffs between different assumptions, and how these assumptions affect the interpretation of study findings.

Box 1. Variance estimator for the non-clustered design

Estimating the model in Equation (2) using OLS produces a difference-in-means estimator, $\hat{\beta}_{ATE} = \bar{y}_T - \bar{y}_C$, with the following variance estimator:

$$\hat{Var}(\hat{\beta}_{ATE}) = \frac{s_T^2}{np} + \frac{s_C^2}{n(1-p)} - \frac{(s_T - s_C)^2}{n},$$

where $s_T^2 = \sum_{i:T_i=1}^{np} (y_i - \bar{y}_T)^2 / (np-1)$ and $s_C^2 = \sum_{i:T_i=0}^{n(1-p)} (y_i - \bar{y}_C)^2 / (n(1-p)-1)$ are sample variances for the treatment and control groups; *n* is the sample size; and *p* is the proportion assigned to the treatment group.

Box 2. Variance estimator for the clustered design

Estimating the model in Equation (3) using weighted least squares produces a weighted difference-in-means estimator, $\hat{\beta}_{ATE,Clus} = (\bar{\bar{y}}_{TW} - \bar{\bar{y}}_{CW})$, where $\bar{\bar{y}}_{TW}$ and $\bar{\bar{y}}_{CW}$ are weighted averages of the cluster means for the treatment and control groups. A variance estimator is as follows:

$$\hat{Var}(\hat{\beta}_{ATE,Clus}) = \frac{s_{TW}^2}{\bar{w}_T^2 mp} + \frac{s_{CW}^2}{\bar{w}_C^2 m(1-p)} - \frac{1}{m}\left(\frac{s_{TW}}{\bar{w}_T} - \frac{s_{CW}}{\bar{w}_C}\right)^2, \text{ where}$$

$$s_{TW}^2 = \frac{1}{mp-1} \sum_{j:T_j=1}^{mp} w_j^2 (\bar{y}_j - \bar{\bar{y}}_{TW})^2, \quad s_{CW}^2 = \frac{1}{m(1-p)-1} \sum_{j:T_j=0}^{m(1-p)} w_j^2 (\bar{y}_j - \bar{\bar{y}}_{CW})^2,$$

$$\bar{w}_T = \frac{1}{mp} \sum_{j:T_j=1}^{mp} w_j, \quad \bar{w}_C = \frac{1}{m(1-p)} \sum_{j:T_j=0}^{m(1-p)} w_j,$$

s_{TW}^2 and s_{CW}^2 are weighted sample variances across clusters; *m* is the number of clusters in the sample; *p* is the proportion of clusters assigned to the treatment group; w_j is the weight assigned to cluster j (for example, the cluster sample size or 1); and $\bar{w}_T$ and $\bar{w}_C$ are average cluster weights.

Box 3. Variance estimator for the non-clustered, random block design

The variance estimator for the weighted difference-in-means estimator across blocks is as follows:

$$\hat{Var}(\hat{\beta}_{ATE}) = \frac{1}{(h-1)h\bar{w}^2} \sum_{b=1}^{h} (w_b \hat{\beta}_{ATE,b} - \bar{w}\hat{\beta}_{ATE})^2,$$

where *h* is the number of blocks; $\hat{\beta}_{ATE,b}$ is the impact estimate in block *b;* w_b is the weight for block *b* (for example, the block sample size or 1); and $\bar{w}$ is the average block weight.

(This page left intentionally blank)

References

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Procedures of the Fifth Berkeley Symposium on Math and Statistical Probability, 1*, 221–233.

Imbens, G. & Rubin, D. (2015). Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge, UK: Cambridge University Press.

Kautz, T., Schochet, P. Z. & Tilley, C. (2017). *Comparing impact findings from design-based and model-based methods: An empirical investigation.* (NCEE 2017-4026). Washington, DC: Analytic Technical Assistance and Development, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Liang, K. & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*, 13–22.

Raudenbush, S. W. & Bryk, A. (2002). Hierarchical linear models: Applications and data analysis methods. Newbury Park, CA: Sage.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Education Psychology, 66*, 688–701.

Rubin, D. B. (1986). Which ifs have causal answers? Discussion of Holland's "Statistics and causal inference." *Journal of the American Statistical Association, 81*, 961–962.

Scher, L. & Cole, R. (2017). *Evidence review standards considerations when using RCT-YES.* Washington, DC: Analytic Technical Assistance and Development, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Accessed at www.rct-yes.com.

Schochet, P. Z. (2016 Second Edition; 2015 First Edition). *Statistical theory for the RCT-YES software: Design-based causal inference for RCTs* (NCEE 2015–4011). Washington, DC: Analytic Technical Assistance and Development, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Accessed at https://ies.ed.gov/ncee/pubs/20154011/pdf/20154011.pdf.

Schochet, P. Z. (2013). Estimators for clustered education RCTs using the Neyman model for causal inference. *Journal of Educational and Behavioral Statistics, 38*(3), 219–238.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*, 817–838.