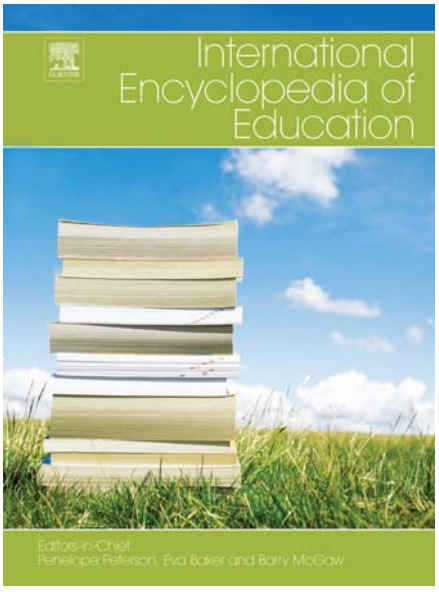

Scheier L M (2010), Methods for Approximating Random Assignment: Regression Discontinuity and Propensity Scores. In: Penelope Peterson, Eva Baker, Barry McGaw, (Editors), International Encyclopedia of Education. volume 3, pp. 104-110. Oxford: Elsevier.

Methods for Approximating Random Assignment: Regression Discontinuity and Propensity Scores

L M Scheier, LARS Research Institute, Las Vegas, NV, USA; Washington University, School of Medicine, St. Louis, MO, USA

Glossary

Bias – A term usually referring to an unwanted or extraneous source of systematic variance affecting the standard error of a treatment effect that arises from several sources (see the definition for the glossary term 'Internal validity'). Usually, minimized or controlled through methods of statistical adjustment or stringent experimental design procedures. Selection bias occurs when there is a different probability of a unit or individual being chosen to participate in a study or assigned to a treatment condition and the characteristics of that individual are confounded with treatment outcomes (e.g., skills, motivation, or preexisting conditions are not independent of the anticipated treatment outcome).

Causal inference – Also called causal propositions, but not the same as causal explanation or causal mechanism, it concerns the expected relationship between some event B that is anticipated to always occur when A happens. The inference or supposition of existence is based on expectation that if all other possible factors that induce a relationship between A and B have been controlled, then the absolute reason for B is always A. In terms of a manipulative account of causation based on experimental work defined as using *prima facie* and, often, probabilistic evidence to infer causal relations (A causes B when the experimenter induces A always).

Confounding – A means of describing relationships between events, where A is the manipulation and B the expected outcome from A. Confounding arises because some measure C is on the path from A to B and mixes its variability so as to "falsely obscure or accentuate the relationship between them" (MacKinnon *et al.*, 2000: 174). When these variances cannot be unambiguously tested, it is usually a sign of confounding. Controlling for the confounder, usually, provides an undistorted estimate of the true relationship between the predictor and outcome in question. Unlike mediators or intervening variables – which carry a causal connotation – confounders are not necessarily posited to cause the linkage between predictor and outcome, (i.e., spuriousness) but are merely related to one or both.

Counterfactual reasoning – A term borrowed from eighteenth-century Scottish philosopher David Hume – a staunch advocate of empiricism in experimental methods – to indicate something that is contrary to fact. In experimental work, a treatment causes an observed effect or alteration on some designated outcome; the counterfactual is what would have happened to the subjects in the absence of a treatment effect (the subjects were designated controls and did not receive the treatment).
A counterfactual inference (based on qualitative distinction) is made experimentally when a treatment condition is pitted against a control group and all other potential factors that might contribute to group differences are made identical between the two groups through statistical control.

Effect – Defined as the difference between what did happen to a treated individual (unit) and what would have happened had the treatment not worked. Requires a counterfactual model in which comparisons exist between a group of individuals exposed to the treatment (observed result) compared to a reasonably similar group held back from receiving the same treatment (unobserved result). A stable phenomenon that scientists seek to explain in terms of behavioral regularities brought about through manipulation.

Internal validity – A system of checks and balances (deductive reasoning) that enables a researcher to make causal inferences about whether a treatment or intervention affects designated outcomes while at the same time eliminating potential sources of bias (i.e., third-variable alternatives). The benchmark criteria vary but have, traditionally, included unambiguous temporal precedence, contemporary history, testing, maturation, pretesting procedures, measurement (reliability), statistical regression, differential selection, mortality, contamination (diffusion of treatment), and compensatory responses (equalization, rivalry, demoralization, etc.).

Propensity score – The probability that an individual is assigned to an experimental treatment group conditional on the individual's covariate information.

Derived as a single scalar function (like a discriminant function) to model a cluster of individual differences variability.

Randomization – A sampling design procedure to ensure that preexisting conditions of the experimental units or observations do not influence variability in the outcomes in question. Will also assure independence of errors at the level of unit of assignment and distribute any possible bias evenly across experimental treatment conditions. Best achieved when subjects are randomly selected, randomly assigned to groups, which are then further assigned randomly to the experimental treatment conditions.

Selection – The process of picking, choosing, or assigning individuals (units) from a larger reservoir to form a homogeneous sample with regard to experimental condition (i.e., treatment vs. control). When done to create two or more probabilistically equivalent groups using strategies of randomization, serves to eliminate bias that might interfere with rendering causal interpretations.

Treatment – A declaration or demarcation of an event manipulation that is oftentimes associated with a structured program, intervention, application, or specific exposure constructed to induce behavioral change defined in terms of a measured endpoint.

Introduction

Randomization is one of many tools researchers have made available to provide experimental control. The English statistician and geneticist Sir Ronald A. Fisher (1935) was one of the first to formally introduce the need for rigorous control in experimental research in order to attribute some element of causation. The emphasis on cause and the need for control arises from a focus on manipulation to induce change in behavior. The term intervention can be used interchangeably with program, treatment, or instructional method. Most common types of interventions in education consist of curricular instruction, programs intended to change behavior (in students, the teacher, or both), improve learning and/or achievement, enrich academic success, and so forth. In psychology, the term intervention is often synonymous with therapeutic engagement (individual or group) or some specified treatment modality (Freudian vs. cognitive-behavioral). In the medical sciences, the term treatment indicates a desired physical change (i.e., reduction in symptoms) through visceral manipulation (removing or adding tissue or bone) or physiological change obtained by invoking a pharmacological drug agent. In some cases, for instance, with psychiatry, symptom reduction can be based on both pharmacological treatment and conjunctive psychotherapy with the goal of symptom amelioration. Regardless of discipline, all references to intervention seem to accentuate a manipulation that is meant to cause something to happen. When this occurs, a behavioral scientist wants to make attributions that the change in behavior was a direct result of some manipulation controlling for other possible sources of variability that may exert an influence. The most common form of a causal statement suggests that some form of a treatment designated as a factor in an experiment results in behavioral change. In its simplest form, the factor is given two levels – one designating experimentally treated participants and the other level connoting those individuals who were not exposed to the treatment. (Conceivably, we could apply the same treatment to the same physical entity requiring only one level; however, there are major drawbacks to this process. Concerns that need to be addressed in this type of design include temporal stability (consideration of constancy of response over time), causal transience (the effect of treatment A does not linger and influence treatment B), and unit homogeneity (all of the physical entities in the study are identical, and thus order of implementation would not be of concern). Were these assumptions valid, the method of differences would apply and randomization not required). For descriptive purposes, the term treated is used to indicate those individuals or experimental units exposed to the intervention (the nomenclature E is often used) and the term control or untreated used to capture individuals assigned to a comparison condition (here C is frequently used). Randomization speaks to the process of how individuals are assigned to the different treatment levels (E and C) and the techniques used to equilibrate both measured and unmeasured pretreatment characteristics during assignment. When a randomized field trial is implemented and all other factors considered (i.e., fidelity of implementation, contamination, and differential attrition, to name a few), it is considered the gold standard for asserting causal inference.

In the case of many educational interventions, a researcher is largely invested in finding out whether a select intervention or program of instruction works by improving academic performance or boosting achievement scores in a designated sample. Here then, the focus is on derivation of a meaningful effect that comes about by contrasting treatment conditions that perform in a specific time frame on a common benchmark. The level of bias in estimating this effect is thus, largely, determined by the method of assignment of individuals or units to the various experimental treatment conditions. All things considered, if participants assigned to either the treatment group (receiving the intervention) or a comparison control group has an equal probability of being assigned to their

respective treatment condition, then the study meets necessary – but not sufficient – conditions to assert some level of causal inference (and can be labeled true experimental or randomized). Should this occur, there is much less concern with selection bias or extraneous confounding nuisance factors that may influence the outcomes in question. One of the major assumptions of randomization is that treatment assignment method is independent of any systematic pretreatment differences that could be used to characterize participants. Maintaining this stringent condition reduces the influence of potential confounds and allows a researcher to make causal assertions about the program of instruction or intervention in question (all factors considered the program or instructional method identified as the treatment is the only variable that can cause the observed differences between the groups). It is easy to see then that the absence of randomization serves as a threat to the internal validity of a study and affects the processes through which scientists make inductive inferences.

This article outlines the requirements for randomization paying particular attention to the underlying philosophical arguments surrounding cause and effect. The article briefly describes the most commonly used methods for randomization, noting what these techniques net in terms of experimental control. In addition, the article offers several alternatives to randomized assignment in the event that such manipulation is not possible or would disturb the scientific evidence. Each strategy is discussed with regard to how it stands up against the gold standard of randomization as well as noting certain pitfalls that might be encountered with inappropriate use.

Purpose of Randomization

Philosophically speaking, one can never really know why something happens nor can one predict events with certainty. To an ever-mindful child, flipping a light switch is the sole reason why the light glows and no other ulterior explanation concerning voltage, impedance, and electric current is required. However, to the more sensitive and knowing adult mind, there is considerable more to the light burning incandescent than the flip of a switch. Even though scientists operate knowing full well that they cannot assert ultimate cause, there is still the desire to learn as much about the world and our human experience as possible. As part and parcel of this epistemological pursuit, scientists engage a logic of causation using an investigative system of reasoning (i.e., inductive, hypothetico-deductive, and abductive) that provides systematic methods to clarify knowledge. This system has undergone considerable change since early Greek philosophers argued about the necessary and sufficient conditions of causation – primarily because modern physical sciences (referring mainly to quantum physics) have cast doubt on whether one will ever observe certain events to occur, but with available mathematical precision will know with certainty (or regularity) they do occur. In an axiomatic system, scientists postulate that things will occur and then through logical deduction gather observational data to strengthen the veracity of original postulates. In subatomic or elementary particle physics, the human mind can rarely truly see events as they unfold, but the same human biocomputing machine is capable of proceeding from axioms and postulates to gather data or evidence of patterns. As a result, the changing veneer of science has stimulated a more modern treatment of the philosophy of causation; one that has fabricated a reasoned stance from which we gather evidence of stable phenomenon.

As philosophical discourse on theories of causation unfolded, philosophers became more engaged in fashioning the particulars of causal inferences and backed off the need to equate causation with explanation. Greater emphasis was placed on the methods scientists use – both in the laboratory as well as the nature of experimental reasoning used to explain their findings. One outgrowth of this emphasis is critical rationalism, which provides epistemic criteria for a scientific method emphasizing not what one knows, but rather how one ascertains what one knows. According to critical rationalism, the goals of science are to rule out false impressions and eliminate inadequate or weak theories. Philosopher Karl Popper – responsible for explicating the tenets of critical rationalism – termed this emphasis on paring away improper explanations the falsification or refutability of a theory. No matter the level of experimental sophistication even that obtained in the hard physical sciences, a theory can never be proved as true; one merely probes the value of our theories as explanations of events. Good theories are the ones that withstand tests attempting to refute them (this has been termed the logic of falsification).

This stance makes all experimental work especially in the behavioral sciences really a statement of probability articulated in terms of how likely something can be attributed to a set of events. As part of the logic of scientific method an investigator frames the null hypothesis (the outcome Y is the same regardless of whether E *vs.* C was administered as a treatment) by a stated conditional significance level to minimize the probability that we would observe this event ($E - C = 0$) by chance alone. Stated differently, if an investigator sets the significance level $p < 0.05$, one could expect the event expressing no difference in treatment conditions (i.e., null hypothesis) to occur 5 times in 100 trials (i.e., chance findings); whereas, the other 95 times, the investigator rejects the null in favor of an alternative hypothesis ($E - C > 0$). The alternative is part of the success of science and occurs because of some systematic reason – likely an effect produced by the treatment in question. This logic of inference helps to frame the

importance of randomization because a researcher wants to avoid chance findings and also eliminate the potential that some systematic but unknown cause was responsible for the desired outcomes.

With this hypothesis-testing framework in mind, a scientist interested in predicting a specific outcome (effect) from an experimental manipulation must rule out that extraneous or unwanted sources of variation (also termed rival hypotheses, alternative explanations, or threats to the interpretation), and also present valid explanations for the outcome – even though these sources of variance were not, directly, part of the manipulation. If the connection between two variables is causal and all other possible contributing events are known, the scientist has taken great strides to ensure the internal validity of the study. If there are plausible threats to internal validity whatever they may be – with random assignment – they are distributed equally over conditions (the extraneous variable is equally like to occur among participants assigned to receive an experimental treatment as in those assigned to the control condition). Based on the principles of randomization, if a researcher were to sample a unit or individual from one level and compare this to a randomly selected unit obtained from another level (here we expressly state that the experimental treatment has two levels), all other factors considered these units would be comparable (this is even more true with repeated samplings). In essence, randomization grapples with differences in people or resources by selecting units or individuals irrespective of any prior existing conditions that may be confounded with the outcome. While randomization cannot offset many plausible and real threats to internal validity, it is a surefire method to eliminate selection bias as a counterfactual position.

Methods of Randomization

Random assignment occurs when the procedure for assigning units or individuals to conditions is based on chance. The most ubiquitous form of randomization involves a simple coin toss. Given the equal probability of obtaining a head or tail from any single independent toss of a fair coin, this represents the simplest way to randomize subjects to experimental treatment conditions. Rolling a fair die can also achieve the same result given the probability of obtaining any number on the face of a die is one-sixth and the probability of obtaining any particular number is governed by chance alone. Simple random sampling can also be accomplished using the lottery method – which involves placing an equal number of distinctly colored small objects into a receptacle of some sort. The process of assignment for simple random sampling involves reaching into the receptacle without looking and removing the objects one by one (without replacement). There is no set or defined way to remove each object; they are just grabbed one by one. After each object is removed, it is assigned to an experimental group in some systematic fashion. For example, the first object pulled out can assign a subject to receive an experimental treatment and the second object pulled out assigns the next subject to the control group with each successive draw assigned in the same prescribed manner. The objects can be colored pieces of paper, paper with numbers scribbled on them that match up to student or patient identifiers, or colored marbles. If a study warrants three experimental conditions with two experimentally treated groups and one comparison control group, then three colors would be used to distinguish assignment procedures. In any of the examples provided above, a simple random number generator using computing resources can be used to create numeric lists. Subjects can then be assigned on the basis of whether their identification number matches the one appearing on a list. In the event that numbers are generated that do not match, these are discarded or ignored. A researcher would move down the list until a match is found and then assign the matched number to experimental condition. To avoid tedium, multiple numbers or ranges of numbers can be used to find suitable matches for personal identifiers (all numbers ending with 1 as in 2001, 3001, and so forth would be assigned to one condition while numbers ending in 2 as in 2562, 3472, and so forth would be assigned to a different condition until all participant's numbers have been exhausted).

In order to truly benefit from random assignment, there must be sufficient numbers of units (individuals) "relative to the variability between units" (Cook and Campbell, 1979: 5). Again, this reinforces the goal of random assignment to equilibrate any potential differences that might exist before a treatment is applied. Designing a process of random assignment around two participants – one male and one female – is unlikely to achieve the desired results, particularly if gender masks other meaningful causal attributes (i.e., sex-linked differences), or has direct relevance to the endpoint. While the need for sufficient sampling speaks directly to the issue of power, it also reinforces the emphasis on average behavior across a representative sampling.

Is Randomization a Panacea?

Despite the overall emphasis on randomization as a means to control extraneous and systematic variance in an experimental framework, there are some concerns that it is not the ultimate panacea. For instance, Cook and Campbell (1979) suggest "it is indeed false to claim that randomization controls for all threats to internal validity" (p. 85). Even with randomization, threats to internal validity can arise from study attrition (i.e., loss of subjects over time), differential attrition (unequal loss of subjects from treatment conditions), and compensatory reactions. The latter

situation, often, occurs if experimental assignment procedures are meant to remove any systematic differences that may result from some privileged resources such as those obtained through status or power (socioeconomically valued factors). In the case of compensatory rivalry, participants in the untreated control condition who do not receive the benefits of an intervention begin to make claims that they too should receive the intervention (and its benefits), and thus disrupt the experiment by creating their own desirable treatment to improve the designated outcomes. Compensatory equalization, often, arises when someone in a position of authority (i.e., management or school administration) decides that participants assigned to the control condition should benefit from an intervention, and thus contaminate the assignment procedure by distributing intervention materials. Compensatory rivalry can be attenuated by promising control participants that they will receive a delayed intervention or providing something of equal value that is unrelated to the target endpoint.

In many instances, randomization is not desirable, feasible, or permissible, and this has led to proliferation of quasi-experimental designs that use various methods to approximate equivalence between experimental units. The nonexperimental group design, interrupted time series design, and other remedies (matching, stratifying, masking or blinding, cutoff-based methods, and regression adjustment with covariates) are used to contrast a treatment condition and comparison group on some outcome of interest. While these remedies afford some element of control, they do not equilibrate preexisting conditions by design, but rather by statistical analysis procedures. Even with these refinements, it is better to actively deliberate about plausible threats to validity before designing a study rather than hoping some laundry list of experimental design procedures would eliminate entirely all threats to validity. The following section explores several commonly used methods that make adjustments for sample selection bias and create the specter of randomization: propensity scores and regression discontinuity (RD).

Propensity Score Method

Propensity scores offer an alternative to randomization with observational studies or quasi-experimental designs when a researcher wants to make causal statements. Without covariate adjustments, a researcher runs the risk of obtaining biased estimates and arriving at erroneous conclusions regarding treatment effectiveness. Proper covariate adjustments for propensity or balancing scores produce unbiased estimates for treatment effects even in the absence of randomization. A propensity score is a derived measure indicating the likelihood of a subject being assigned to the treatment group based solely on that subject's covariate information. In the case of two experimental conditions, where individuals are assigned to either receive an intervention or assigned to some type of control comparison group, the propensity score method replaces the confounding covariate measures with a single coarse function that represents the conditional probability of treatment assignment. Derivation of a conditional probability (scalar propensity) score helps adjust any differences between groups (produces an unbiased estimate of average treatment effect) based on a known set of observed covariates (i.e., pretreatment measures).

In many cases, stratification (i.e., subclassification) of the sample based on at most a few covariates would be ideal. Comparisons could then be made between the different cells with the assumption that treatment and control participants are equally represented within each cell. However, oftentimes, the number of possible confounding covariates is large, leading to the likelihood of sparse cells containing either only treatment or control participants. This would truncate the possibility of estimating a treatment effect within the particular strata. In addition, proliferation of strata from excessive covariates creates problems with cross-classification and the inability to obtain proper diagnostics on joint distributions of regressors. When this occurs, propensity score methods provide an ideal means to achieve parsimonious representation of all the different observed or known characteristics that can differentiate treatment and control participants (i.e., this multivariate difference is akin to a linear discriminant function).

Under the stable unit-treatment value assumption, scalar propensity scores are akin to a missing data problem and will satisfy this condition if the treatment assignment is strongly ignorable given the balancing score (the scalar propensity based on a vector of covariates) – that is, the relationship between treatment assignment and the designated endpoint are conditionally independent of the covariates. With a propensity score adjustment, everyone in the population has an equal chance to be assigned to either the treatment or control group, irrespective of any predisposing characteristics (covariates). Once propensity scores are created – and using the *t*-statistic as a gauge – participants can then be iteratively balanced into homogeneous bins or subclasses based on the distribution of estimated propensity scores (five subclasses or quintiles appear sufficient). The end result is that individuals located within a bin have equal probabilities of being selected into either experimental condition (treatment vs. control), and can be considered to have been randomly assigned to their respective conditions. (The propensity score method is based on the notion of equality of variance/covariances within subclasses. If balancing is not achieved, then the model must be reformulated so that the likelihood of being assigned to treatment or control based on propensity scores is equal within a subclass.) Estimates of treatment effects can then be generated within each strata

(quintile) and averaged over the number of strata created. This approach works best with large samples given the expected distributional balance on the covariates in different strata or subclasses. With smaller sample sizes, the anticipated balance may come into question as would be expected with any sample shrinkage and corresponding inflation of standard errors.

Regression Discontinuity

The term regression discontinuity (RD) connotes a type of quasi-experimental design that relies on standard regression methods to obtain unbiased causal estimates of a treatment effect. In order to create boundaries for delineating assignment to experimental condition, an investigator will create a cutoff point using an assignment variable. The assignment variable cannot be caused by the treatment, but does not have to be a pretest score equivalent in form to the posttest (a posttest score is required). It is best – as a means of increasing power – that the assignment variable has minimal overlap with treatment and often multiple assignment measures in the form of an index will improve power. The sample is then divided into two components based on this cutoff score with those scoring on one side of the cutoff assigned to one experimental condition and those scoring on the other side assigned to a different condition. For example, students scoring below the cutoff point on the assignment variable (and, perhaps, showing a deficit in performance) can be designated to participate in a remedial program emphasizing reading or mathematical skills, or some type of instructional curriculum that can enhance their academic aptitude. Those students scoring above the cutoff point are considered the control or comparison group. A scatterplot depicting the assignment scores plotted on the *X* or horizontal axis and the posttest scores on the *Y* or vertical axis might show a vertical displacement at the cutoff score if the remedial program has an effect (i.e., the treatment improves participating students' performance), whereas no change or displacement would be expected if there was no effect. The displacement can be witnessed as either a shift in means reflecting treatment benefits measured in units of the outcome or a change in slope of the regression line at the designated cutoff point. Shifts – both in means or slopes – give the name RD and the size or magnitude of the discontinuity is the size of the effect resulting from the treatment. When there is no program or treatment effect, the functional forms of the slopes will be equivalent – holding all other factors constant.

Greater confidence is gained that the remedial program was responsible for the overall treatment effect if all students below (or above) the cutoff benefit equally from the treatment. It is essential to rule out model misspecification (nonlinearity) and to ensure the performance shift in the posttest scores is clearly demarcated by the cutoff; that is, the same treatment benefit is not observed at other random places along the plot that could support rival interpretations. Moreover, rival hypotheses would have to achieve the precise same effect, using the same markers of performance, and under the same conditions without the benefit of the cutoff. Possible threats to internal validity would have to cause a discontinuity at the precise point of the cutoff and this is highly unlikely, although not implausible. Attrition (from one side of the cutoff) could reduce power and bias the sample as can history, which can influence participants on one side of the cutoff at the exclusion of others. One caution to using this method is the effect of restricted range on variance estimates commonly known as shrinkage. The way to offset shrinkage involves the use of multiple discontinuities to retain the original variances based on a broader range of values for the predictor measure. It is also unlikely that the students in the two boundary conditions would be matched on any aggregate characteristics as encountered with propensity score methods using observed measures. These gray areas make it unlikely that this method will produce efficient and stable estimates of treatment effectiveness without large samples, precise and meaningful cutoff points for assignment, and efforts at cross-validation.

Summary

Randomization affords researchers a clear method to ensure that any differences between units or individuals assigned to experimental conditions are by chance alone. It is an optimal design strategy to fend off certain threats to internal validity and provides the foundation for true experiments seeking to make valid causal inferences. When subjects have an equal probability of being selected and assigned to an experimental condition the unique characteristics they bring to the laboratory or field experiment are equated across treatment conditions. Simply put, the goal of randomization is to seek a level of equality of subjects prior to their assignment to experimental condition. When this occurs, researchers can then make more confident assertions whether a specific manipulation results in the anticipated treatment effect by design as opposed to happening by chance alone.

In the realm of educational studies, it is not always possible to assign students (or teachers) to treatment conditions using random assignment methods. In many cases, well-known studies examining the role of school vouchers, private versus public school education, efficacy of school-based drug-prevention, grade retention, and evaluations of many remedial instructional modalities to improve learning and achievement take shape as observational, quasi-experimental studies lacking the precision afforded by random assignment. These studies tend to be more economical and less cumbersome than randomized trials;

however, the absence of complete randomization hampers scientists' abilities to make valid causal inferences. In recent years, several alternative approaches have been proposed to accommodate the necessity of controlling nuisance factors that may diminish the authority of causal attributions. With these tools in hand, a researcher is much closer to being able to state that a certain manipulation resulted in a specific effect and thus reinforce the cause–effect relationship that is the backbone of all scientific effort.

Although several viable alternatives to randomization exist, two in particular are covered in this article (instrumental variables and fixed-effect methods as possible remedies to randomization are discussed elsewhere in the encyclopedia). Propensity scores provide a parsimonious and efficient remedy to the problem of obtaining unbiased estimates of treatment effectiveness by adjusting or balancing treatment group differences based on a single composite characteristic. Any bias associated with treatment condition assignment is controlled statistically through the covariate adjustment, and subjects are balanced on their propensity for selection. Importantly, models using this approach are only as valid as the model selection process used to include covariates in the scalar function. Hidden or missing covariates can differentiate participants in ways not considered and alter the statistical outcomes or at the very least undermine confidence in any causal interpretation.

Another remedy discussed in this article involves dichotomization of samples in a way that permits investigators to make meaningful comparisons or treatment contrasts as if the subjects had been randomly assigned to discrete experimental groups. While these variants on randomization have less stringent requirements for balancing preexisting differences, they still afford a means of comparison not available with subclassification or other covariate-adjustment methods. The technique of RD works on the assumption that oftentimes natural boundary conditions mimic random assignment to treatment conditions. With RD methods, an investigator formulates the designation (assignment) of experimental versus control groups based on students achieving just above (or below) a certain threshold on some benchmark performance criteria. As a special feature, the selection mechanism using the assignment variable cutoff point is fully known (i.e., there is no measurement error), and there is no hidden bias that can influence the estimate of the treatment effect. The elegance of this approach is that it can be mixed with randomization techniques, utilize multiple cutoff points, and incorporate more than one treatment. When certain assumptions regarding internal validity are met, the resulting estimate of treatment effect is virtually unbiased.

All told, despite differences in these two popular methods to approximate randomization, both are similarly geared toward removing extraneous variance or controlling for preexisting treatment group differences that might bias the estimation of treatment effects. Random assignment is a major canon of experimental design but not the *sine qua non* that defines experimentation. Other important components include selection of the treatment, dosage, method of dispensing treatment, measuring and timing of effects, choosing participants, the nature of comparison, and assignment protocols. Even with these additional requirements, the goal of any experimental design or sampling strategy is still to ensure that causal assertions are not mistaken, implausible, or falsifiable.

Bibliography

Cook, T. D. and Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston, MA: Houghton Mifflin.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

MacKinnon, D. P., Krull, J. L., and Lockwood, C. M. (2000). Equivalence of the mediation, confounding, and suppression effect. *Prevention Science* **1**, 173–181.

Further Reading

Blackwell, E. and Hodges, J. L. (1957). Design for the control of selection bias. *Annual Mathematical Statistics* **28**, 449–460.

Chalmers, A. (2002). Experiment and the growth of experimental knowledge. In Gärdenfors, P., Wolenski, J., and Kijania-Placek, K. (eds.) *In the Scope of Logic, Methodology and Philosophy of Science*, pp 157–169. Dordrecht: Kluwer Academic Publishers.

D'Agostino, R. B. (1998). Propensity score methods for bias reduction in the comparison of treatment to a non-randomized control group. *Statistics in Medicine* **17**, 2265–2281.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58**, 403–417.

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* **69**, 201–209.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrika* **71**, 1161–1189.

Indurkhya, A., Mitra, N., and Scrag, D. (2006). Using propensity scores to estimate the cost-effectiveness of medical therapies. *Statistics in Medicine* **25**, 1561–1576.

Jacob, B. and Lefgren, L. (2004). Remedial education and student achievement: A regression discontinuity analysis. *Review of Economics and Statistics* **86**, 226–244.

Popper, K. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Routledge Classics.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustments to control bias in observational studies. *Journal of the American Statistical Association* **74**, 318–324.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* **127**, 757–763.

Rubin, D. B. and Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics* **52**, 249–264.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Casual Inference*. New York: Houghton Mifflin.

Sobel, M. E. (1995). Causal inference in the social and behavioral sciences. In Arminger, G., Clogg, C. C., and Sobel, M. E. (eds.) *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, pp 1–38. New York: Plenum Press.