Causal Inference

D B Rubin, Harvard University, Cambridge, MA, USA

A Framework for Causal Inference – Basic Building Blocks

The framework for causal inference that is discussed here is now commonly referred to as the Rubin Causal Model (RCM; Holland, 1986), for a series of articles written in the 1970s (Rubin, 1974, 1976, 1977, 1978, 1980). Other approaches to causal inference, such as graphical ones (e.g., Pearl, 2000), are conceptually less satisfying, for reasons discussed, for instance, in Rubin (2004b, 2005). The presentation here is essentially a brief and relatively nontechnical version of that given in Rubin (2006).

For causal inference, there are several basic building blocks. A unit is a physical object, for example, a person, at a particular point in time. A treatment is an action that can be applied or withheld from that unit. We focus on the case of two treatments, although the extension to more than two treatments is simple in principle but not necessarily so with real data. Associated with each unit are two potential outcomes: the value of an outcome variable Y (e.g., test score) at a point in time t when the active treatment (e.g., new educational program) is used at an earlier time t_0, and the value of Y at time t when the control educational program is used at t_0. The objective is to learn about the causal effect of the application of the active treatment relative to the control (treatment) on Y. Formal notation for this meaning of a causal effect first appeared in Neyman (1923) in the context of randomization-based inference in randomized experiments. Let W indicate which treatment the unit received: $W = 1$ the active treatment, $W = 0$ the control treatment. Moreover, let $Y(1)$ be the value of Y if the unit received the active version, and $Y(0)$ the value if the unit received the control version. The causal effect of the active treatment relative to its control version is the comparison of $Y(1)$ and $Y(0)$ – typically the difference, $Y(1) - Y(0)$, or perhaps the difference in logs, $\log[Y(1)] - \log[Y(0)]$, or some other comparison, possibly the ratio. The fundamental problem for causal inference is that, for any individual unit, we can observe only one of $Y(1)$ or $Y(0)$, as indicated by W; that is, we observe the value of the potential outcome under only one of the possible treatments, namely the treatment actually assigned, and the potential outcome under the other treatment is missing. Thus, inference for causal effects is a missing-data problem – the "other" value is missing. Of importance in educational research, the gain score for a unit, posttest minus pretest, measures a change in time, and so is not a causal effect.

We learn about causal effects using replication, which involves the use of more than one unit. The way we personally learn from our own experience is replication involving the same physical object (me or you) with more units in time, thereby having some observations of $Y(0)$ and some of $Y(1)$. When we want to generalize to units other than ourselves, we typically use more objects; that is what is done in social science experiments, for example, involving students and possible educational interventions, such as value-added assessment (e.g., Rubin *et al.*, 2004). Replication does not help without additional assumptions. The most straightforward assumption to make is the stable unit treatment value assumption (SUTVA; Rubin, 1980, 1990) under which the potential outcomes for the ith unit are determined by the treatment the ith unit received. That is, there is no interference between units (Cox, 1958) and there are no versions of treatments (Rubin, 1980). Then, all potential outcomes for N units with two possible treatments can be represented by an array with N rows and two columns, the ith unit having a row with two potential outcomes, $Y_i(0)$ and $Y_i(1)$, where each could, in principal, be a vector with many components. Obviously, SUTVA is a major assumption. Good researchers attempt to make such assumptions plausible by the design of their studies. For example, SUTVA becomes more plausible when units are isolated from each other, as when using, for the units, intact schools rather than students in the schools when studying an educational intervention, such as a smoking prevention program (e.g., see Peterson *et al.*, 2000).

In addition to (1) the vector indicator of treatments for each unit in the study, $W = \{W_i\}$, (2) the array of potential outcomes when exposed to the active treatment, $Y(1) = \{Y_i(1)\}$, and (3) the array of potential outcomes when not exposed, $Y(0) = \{Y_i(0)\}$, we have (4) an array of covariates $X = \{X_i\}$, which are, by definition, unaffected by treatment, such as age, race sex, or pretest scores, where the 'pre' means prior to the intervention, that is, before t_0. Covariates can be used to help define causal estimands. All causal estimands involve comparisons of $Y_i(0)$ and $Y_i(1)$ on either all N units, or a common subset of units; for example, the average causal effect across all units that are female as indicated by their X_i, or the median causal effect for units with X_i indicating male and $Y_i(0)$ indicating failure on the posttest under the control treatment.

Under SUTVA, all causal estimands can be calculated from the matrix of scientific values with ith row: $(X_i, Y_i(1), Y_i(0))$. By definition, all relevant information is encoded in X_i, $Y_i(0)$, $Y_i(1)$ and so the labeling of the N rows is a random permutation of $1, \ldots, N$, and the matrix is row exchangeable. Covariates play a particularly important role in the analysis of observational studies for causal effects where they are also known as possible confounders or risk factors. In some studies, the units exposed to the active treatment differ in their distribution of covariates in important ways from the units not exposed. To see how this issue influences our formal framework, we must define the assignment mechanism, the probabilistic mechanism that determines which units receive the active version of the treatment and which units receive the control version.

The Assignment Mechanism – Motivating Examples

Even with SUTVA, inference for causal effects requires the specification of an assignment mechanism: a probabilistic model for how some units were selected to receive the active treatment and how other units were selected to receive the control treatment. We first illustrate this model in two trivial artificial examples, and then present formal notation for this model.

Consider a teacher who is considering one of two treatments to apply to each of eight students, a standard and a new one. This teacher is a great teacher and the treatment that is best for each student! When they are equally effective, a fair coin is tossed. **Table 1** gives both the hypothetical potential outcomes in test scores under each treatment for these students, and also their individual causal effects. The column labeled W shows which treatment each student received, $W_i = 0$ or $W_i = 1$ for the ith student.

Notice that the eight individual causal effects indicate that the typical student will do better with the standard treatment (i.e., the control treatment): the average causal effect is two points in favor of the standard. However, the teacher, who is conducting ideal educational practice for the benefit of the students, reaches the opposite conclusion from an examination of the observed data: the students assigned the new treatment do, on average more than twice as well as the students assigned the control, with absolutely no overlap in their distributions!

Table 1 Perfect teacher

Potential outcomes			*Observed data*		
	$Y(0)$	$Y(1)$	W	$Y(0)$	$Y(1)$
	13	14	1	?	14
	6	0	0	6	?
	4	1	0	4	?
	5	2	0	5	?
	6	3	0	6	?
	6	1	0	6	?
	8	10	1	?	10
	8	9	1	?	0
True averages	7	5	Observed averages	5.4	11

What is wrong? The simple comparison of observed outcomes assumes that treatments were randomly assigned, rather than as they were, to provide maximal benefit to the students. More precisely, notice that the teacher, by comparing observed means of the outcome Y, is using the three observed values of $Y_i(1)$ to represent the five missing values of $Y_i(1)$, effectively imputing or filling in $\overline{y}_1$, the observed mean of the $Y_i(1)$, for the five $Y_i(1)$ question marks, and analogously effectively filling in $\overline{y}_0$ for the three $Y_i(0)$ question marks. This process makes sense for point estimation if the three observed values of $Y_i(1)$ were randomly chosen from the eight values of $Y_i(1)$, and the five observed values of $Y_i(0)$ were randomly chosen from the eight values of $Y_i(0)$. However, under the actual assignment mechanism, it does not make sense. It would make much more sense under the actual assignment mechanism to impute the missing potential outcome for each student to be less than or equal to that student's observed potential outcome. The point here is simply that the assignment mechanism is crucial to valid inference about causal effects, and the teacher used a nonignorable assignment mechanism (defined shortly). With a posited assignment mechanism, it is possible to draw causal inferences; without one, it is impossible. It is in this sense that when drawing causal inferences, a model for the assignment mechanism is more fundamental than a model for the potential outcomes.

We next consider a classical paradox in educational research that is easily resolved with the simple ideas we have already presented, despite the controversy that the paradox engendered in some literatures. This example illustrates how important it is to keep this perspective clearly in mind when thinking about causal effects of interventions. Lord (1967) proposed the following example:

> A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex differences in these effects. Various types of data are gathered. In particular, the weight of each student at the time of arrival in September and the following June are recorded.

The result of the study for the males is that their average weight is identical at the end of the school year to what it was at the beginning; in fact, the whole distribution of weights is unchanged, although some males lost weight and some males gained weight – the gains and losses exactly balance. The same thing is true for

the females. The only difference is that the females started and ended the year lighter on average than the males. On average, there is no weight gain or weight loss for either males or females. From Lord's description of the problem quoted above, the quantity to be estimated, the estimand, is the difference between the causal effect of the university diet on males and the causal effect of the university diet on females. That is, the causal estimand is the difference between the causal effects for males and females, the differential causal effect.

The paradox is generated by considering the contradictory conclusions of two statisticians. Statistician 1 observes that there are no differences between the September and June weight distributions for either males or females. Thus, statistician 1 concludes that

> . . .as far as these data are concerned, there is no evidence of any interesting effect of diet (or of anything else) on student weight. In particular, there is no evidence of any differential effect on the two sexes, since neither group shows any systematic change (Lord, 1967: 305).

Statistician 2 looks at the data in a more sophisticated way. Effectively, he examines males and females with about the same initial weight in September, say a subgroup of overweight females (meaning simply above-average-weight females) and a subgroup of underweight males (analogously defined). He notices that these males tended to gain weight on average and these females tended to lose weight on average. He also notices that this result is true no matter what group of initial weights he focuses on. (Actually, Lord's statistician 2 used a covariance adjustment/regression adjustment.) Therefore the conclusion of statistician 2 is that, after controlling for initial weight, the diet has a differential positive effect on males relative to females because for males and females with the same initial weight, on average, the males gain more than the females.

Let us formulate the problem as advocated here. The units are the students, the time of application of treatment (the university diet) is September, and the time of the recording of the outcome Y is June; accept the stability assumption. The potential outcomes are June weight under the university diet $Y_i(1)$ and under the control diet $Y_i(0)$. The covariates are sex of students, male versus female, and September weight. However, the assignment mechanism has assigned everyone to the new treatment! There is no one, male or female, who is assigned to the control treatment. Hence, there is absolutely no purely empirical basis on which to estimate the causal effects, either raw or differential, of the university diet relative to the control diet. Lord has created partial confusion by making the problem complicated with the introduction of the covariates male/female and initial weight, and by using the observed variable notation, Y_{obs}, where $Y_{obs,i} = W_i\ Y_i(1) + (1-W_i)(Y_i(0))$, which mixes up the science, reflected here by $(Y_i(1), Y_i(0))$ and the assignment mechanism, reflected by W_i. For more statistical details of the resolution of this paradox, see Holland and Rubin (1983), and for earlier related discussion, see for example, Lindley and Novick (1981), or Cox and McCullagh (1982). However, the point here is that the paradox is immediately resolved through the explicit use of potential outcomes. In fact, either answer could be correct for causal inference depending on what we are willing to assume about the control diet, as elucidated by Holland and Rubin (1983).

The Assignment Mechanism – Formal Notation

A model for the assignment mechanism is needed for all forms of statistical inference for causal effects. Formally, the assignment mechanism gives the conditional probability of each vector of assignments given the covariates and potential outcomes:

$$\Pr(W|X, Y(0), Y(1)).$$

A specific example of an assignment mechanism is a completely randomized experiment with N units, where $n < N$ are assigned to the active treatment, and $N - n$ to the control treatment.

$$\Pr(W|X, Y(0), Y(1)) = \begin{array}{ll} 1/C_n^N & \text{if } \Sigma W_i = n \\ 0 & \text{otherwise} \end{array}$$

An unconfounded assignment mechanism is free of dependence on either $Y(0)$ or $Y(1)$:

$$\Pr(W|X, Y(0), Y(1)) = \Pr(W|X).$$

With an unconfounded assignment mechanism, at each set of values of X_i that has a distinct probability of $W_i = 1$, there is effectively a completely randomized experiment. That is, if X_i indicates sex, with males having probability 0.2 of receiving the active treatment and females having probability 0.5 of receiving the active treatment, then essentially one randomized experiment is prescribed for males and another for females.

The assignment mechanism is probabilistic if each unit has a positive probability of receiving either treatment:

$$0<\Pr(W_i = 1|X, Y(0), Y(1))<1,$$

where the unit level probabilities are known as propensity scores (Rosenbaum and Rubin, 1983). Unconfounded probabilistic assignment mechanisms often allow particularly straightforward estimation of causal effects from all perspectives, and these assignment mechanisms form the basis for inference for causal effects in more complicated situations, such as when assignment probabilities depend on covariates in unknown ways, or when there is noncompliance with the assigned treatment, or even

in observational (nonrandomized) studies. Unconfounded probabilistic assignment mechanisms are essentially generalized randomized experiments, and are called strongly ignorable (Rosenbaum and Rubin, 1983).

A confounded assignment mechanism is one that depends on the potential outcomes. A special class of possibly confounded assignment mechanisms is particularly important to Bayesian inference: ignorable assignment mechanisms (Rubin, 1978). Ignorable assignment mechanisms are defined by their freedom from dependence on any missing potential outcomes:

$$\Pr(W|X, Y(0), Y(1)) = \Pr(W|X, Y_{\text{obs}}).$$

Ignorable but confounded assignment mechanisms arise in practice, most commonly in sequential experiments, where the next (in time) unit's probability of being exposed to the active treatment depends on the success rate of those previously exposed to the active treatment versus the success rate of those exposed to the control treatment, as in play-the-winner designs (e.g., Efron, 1971). All unconfounded assignment mechanisms are ignorable, but not all ignorable assignment mechanisms are unconfounded (e.g., play-the-winner designs).

Modes of Causal Inference

There are two distinct forms of assignment-mechanism-based (or randomization-based) modes of causal inference: one due to Neyman (1923) and the other due to Fisher (1925). There is a third approach (Rubin, 1978), which is posterior predictive (Bayesian).

Fisher's approach is closely related to the mathematical idea of proof by contradiction. The first element in Fisher's mode is the null hypothesis, which is usually $Y_i(1) \equiv Y_i(0)$ for all units: the treatments have absolutely no effect on the potential outcomes. Under this null hypothesis, all potential outcomes are known from the observed values of the potential outcomes, Y_{obs}, because $Y(1) \equiv Y(0) \equiv Y_{\text{obs}}$. It follows that, under this null hypothesis, the value of any statistic, S, such as the difference of the observed averages for units exposed to treatment 1 and units exposed to treatment 0, $\overline{y}_1 - \overline{y}_0$, is known, not only for the observed assignment, but also for all possible assignments W. Suppose we calculate the value of S under each possible assignment (assuming the null hypothesis) and also calculate the probability of each assignment under the randomized assignment mechanism. Knowing the value of S for each W and its probability, we can then calculate the probability (under the assignment mechanism and the null hypothesis) that we would observe a value of S as unusual as, or more unusual than, the observed value of S, S_{obs}. Unusual is defined *a priori*, typically by how discrepant S_{obs} is from the typical values of S. This probability is the plausibility (p-value or significance level) of the observed value of the statistic S under the null hypothesis: if the null hypothesis were true, the probability of S being as rare, or more rare, than S_{obs}.

Neyman's form of randomization-based inference can be viewed as drawing inferences by evaluating the expectations of statistics over the distribution induced by the assignment mechanism in order to calculate a confidence interval for the typical causal effect. First, an unbiased estimator of the causal estimand (the typical causal effect, e.g., the average) is created, and an unbiased, or upwardly biased, estimator of the sampling variance of that unbiased estimator is found (bias and sampling variance both defined with respect to the randomization distribution). Then, an appeal is made to the central limit theorem for the normality of the estimator over its randomization distribution, whence a confidence interval for the causal estimand is obtained.

With a data set that is not from a randomized study, we try to structure the problem so that we can conceptualize the data as having arisen from an underlying randomized experiment, and then estimate the assignment mechanism via the propensity scores for all the units. A key idea is that, like good experiments, good observational studies are designed, not simply found (Rubin, 2002, 2008). When designing an experiment, we do not have any outcome data, but we plan the collection, organization, and analysis of the data to improve our chances of obtaining valid, reliable, and precise causal answers. The same exercise should be done in an observational study: even if outcome data are available at the design stage, they should be set aside. As observational studies are rarely known to be unconfounded, we are concerned with the sensitivity of answers to unobserved covariates. The methods described by Rosenbaum (2002) are appropriate from the randomization-based perspective.

Posterior predictive (Bayesian) causal inference for causal effects requires a model for the underlying data, $\Pr(X, Y(0), Y(1))$, and this is where science enters. A virtue of the RCM framework is that it separates science – a model for the underlying data, from what we do to learn about science – the assignment mechanism, $\Pr(W|X, Y(0), Y(1))$. This approach directly and explicitly confronts the missing potential outcomes, $Y_{\text{mis}} = \{Y_{\text{mis},i}\}$, where $Y_{\text{mis},i} = W_i Y_i(0) + (1 - W_i) Y_i(1)$. The perspective takes the specification for the assignment mechanism and the specification for the underlying data, and derives the posterior predictive distribution of Y_{mis}, that is, the distribution of Y_{mis} given all observed values:

$$\Pr(Y_{\text{mis}}|X, Y_{\text{obs}}, W).$$

From this distribution and the observed values of the potential outcomes, the observed assignments, and observed covariates, the posterior distribution of any causal effect can,

in principle, be obtained. This conclusion is immediate if we view the above posterior predictive distribution as specifying how to take a random draw of Y_{mis}. Once a value of Y_{mis} is drawn, any causal effect can be directly calculated from the drawn value of Y_{mis} and the observed values of X and Y_{obs}. Repeatedly drawing values of Y_{mis} and calculating the causal effect for each draw generates the posterior distribution of the desired causal effect. Thus, as in Rubin (1978), we can view causal inference entirely as a missing data problem, where we multiply impute (Rubin, 1987, 2004a) the missing potential outcomes to generate a posterior distribution for the causal effects.

With ignorable treatment assignment, all that we need model is the science $\Pr(X, Y(0), Y(1))$, and we can ignore the assignment mechanism. A strength of this model-based approach is that it allows us to conduct causal inference by predicting the missing potential outcomes from observed values. The problem with this approach is the need to specify the distribution $\Pr(X, Y(0), Y(1))$, which sometimes can implicitly involve extrapolations that are extremely unreliable. More details of this approach are beyond the scope of this article, but can be found in, for example, Rubin (2006). With nonignorable treatment assignment, the simplifications previously described do not follow in general, and the analysis typically becomes far more difficult and speculative.

Complications

There are many complications that occur in real-world studies for causal effects, many of which can be handled much more flexibly with the Bayesian approach than with randomization-based methods. Of course, the models involved, including associated prior distributions, can be very demanding to formulate in a practically reliable manner. In addition, Neymanian evaluations are still important.

Most of the field of classical experiment design is devoted to issues that arise with more than two treatment conditions (e.g., Kempthorne, 1952; Cochran and Cox, 1957).

Missing data, due perhaps to unit dropout or machine failure, can complicate analyses more than one would expect based on a cursory examination of the problem. Methods such as multiple imputation (Rubin, 1987, 2004a), the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977), data augmentation (Tanner and Wong, 1987), and the Gibbs sampler (Geman and Geman, 1984) are fully compatible with the Bayesian approach to causal inference outlined in the section titled 'Modes of causal inference'. Gelman *et al.* (2003) provide guidance on many of these issues from the Bayesian perspective.

Another complication, common when the units are people, is noncompliance (e.g., as in Sommer and Zeger, 1991). Early work related to this issue can be found in economics (e.g., Tinbergen, 1930; Haavelmo, 1944) and elsewhere (e.g., Zelen, 1979; Bloom, 1984). Much progress has been made in recent years on this topic (e.g., Baker, 1998; Baker and Lindeman, 1994; Goetghebeur and Molenberghs, 1996; Angrist *et al.*, 1996; Imbens and Rubin, 1997; Little and Yau, 1998; Hirano *et al.*, 2000).

Further complications include truncation due to death (Rubin, 2000, 2004b; Zhang and Rubin, 2003; Zhang *et al.*, 2007). In the real world, complications typically do not appear simply one at a time. For example, a randomized experiment in education evaluating school choice suffered from missing data in both covariates and longitudinal outcomes; besides, the outcome was multicomponent at each point in time; in addition, it suffered from noncompliance that took several levels because of the years of school, as discussed, for example, Barnard *et al.* (2003) in the context of the school choice example.

Many of the above complications can be viewed as special cases of principal stratification (Frangakis and Rubin, 2002). This appears to be an extremely fertile area for research and application of Bayesian methods for causal inference, especially using modern simulation methods (see, e.g., Gilks *et al.*, 1995).

See also: Analysis of Covariance; Instrumental Variables; Methods for Approximating Random Assignment; Missing Data; Multivariate Linear Regression; Observational Studies; Quasi-Experimentation: Two Group Design.

Bibliography

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91**, 444–472.

Baker, S. G. (1998). Analysis of survival data from a randomized trial with all-or-none compliance: Estimating the cost-effectiveness of a cancer screening program. *Journal of the American Statistical Association* **93**, 929–934.

Baker, S. G. and Lindeman, K. S. (1994). The paired availability design: A proposal for evaluating epidural analgesia during labor. *Statistics in Medicine* **13**, 2269–2278.

Barnard, J., Hill, J., Frangakis, C., and Rubin, D. (2003). School choice in NY city: A Bayesian analysis of an imperfect randomized experiment. In Gatsonis, C., Carlin, B., and Carriquiry, A. (eds.) *Case Studies in Bayesian Statistics,* vol. V, pp 3–97. New York: Springer.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review* **8**, 225–246.

Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs*, 2nd edn. (reprinted as a ''Wiley Classic'' (1992)) New York: Wiley.

Cox, D. R. (1958). *The Planning of Experiments*. New York: Wiley.

Cox, D. R. and McCullagh, P. (1982). Some aspects of covariance. *Biometrics* **38**, 541–561.

Dempster, A. P., Laird, N., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

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

Fisher, R. A. (1925). *Statistical Methods for Research Workers*, 1st edn. Edinburgh: Oliver and Boyd.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21–29.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis*, 2nd edn. New York: CRC Press.
Geman, S. and Geman, D. (1984). Stochastic relaxation. *Gibbs Distributions, and the Bayesian Restoration of Images, IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.
Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1995). *Markov Chain Monte Carlo in Practice*. New York: CRC Press.
Goetghebeur, E. and Molenberghs, G. (1996). Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. *Journal of the American Statistical Association* **91**, 928–934.
Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica* **15**, 413–419.
Hirano, K., Imbens, G., Rubin, D. B., and Zhou, X. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1**, 69–88.
Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* **81**, 945–970.
Holland, P. W. and Rubin, D. B. (1983). On Lord's paradox. In Wainer, S. and Messick, H. (eds.) *Principals of Modern Psychological Measurement: A Festschrift for Frederic M. Lord*, pp 3–25. Hillsdale, NJ: Earlbaum.
Imbens, G. W. and Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics* **25**, 305–327.
Kempthorne, O. (1952). *The Design and Analysis of Experiments*. New York: Wiley.
Lindley, D. V. and Novick, M. R. (1981). The role of exchangeabiltiy in inference. *Annals of Statistics* **9**, 45–58.
Little, R. J. and Yau, L. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods* **3**, 147–159.
Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin* **68**, 304–305.
Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translated in *Statistical Science* (1990), 465–472.
Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge, UK: Cambridge University Press.
Peterson, A. V., Jr., Kealey, K. A., Mann, S. L., Marek, P. M., and Sarason, I. G. (2000). Hutchinson smoking prevention project: Long-term randomized trial in school-based tobacco use prevention-results on smoking. *Journal of the National Cancer Institute* **92**, 1979–1991.
Rosenbaum, P. R. (2002). *Observational Studies*, 2nd edn. New York: Springer.
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. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.
Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.
Rubin, D. B. (1977). Assignment of treatment group on the basis of a covariate. *Journal of Educational Statistics* **2**, 1–26.
Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics* **7**, 34–58.
Rubin, D. B. (1980). Comment on ''randomization analysis of experimental data: The Fisher randomization test'' by D. Basu. *Journal of the American Statistical Association* **75**, 591–593.
Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science* **5**, 472–480.
Rubin, D. B. (2000). The utility of counterfactuals for causal inference. Comment on A. P. Dawid, ''Causal inference without counterfactuals.'' *Journal of the American Statistical Association* **95**, 435–438.
Rubin, D. B. (2002). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology* **2**, 169–188.
Rubin, D. B. (2004a). *Multiple Imputation for Nonresponse in Surveys*. (reprinted with new appendices as a ''Wiley Classic.''). New York: Wiley.
Rubin, D. B. (2004b). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* **31**, 161–170; 195–198.
Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. 2004 Fisher lecture. *Journal of the American Statistical Association* **100**, 322–331.
Rubin, D. B. (2006). Statistical inference for causal effects, with emphasis on applications in psychometrics and education. In Rao, C. R. and Sinharay, S. (eds.) *Handbook of Statistics, Volume 26: Psychometrics*, pp 769–800. Amsterdam: Elsevier.
Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics* **2**(3), 808–840.
Rubin, D. B., Stuart, E. A., and Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics* **29**, 103–116.
Sommer, A. and Zeger, S. (1991). On estimating efficacy from clinical trials. *Statistics in Medicine* **10**, 45–52.
Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**, 528–550.
Tinbergen, J. (1930). Determination and interpretation of supply curves: An example. (reprinted in Henry and Morgan (eds.) *The Foundations of Economics) Zeitschrift fur Nationalokonomie.*
Zelen, M. (1979). A new design for randomized clinical trials. *New England Journal of Medicine* **300**, 1242–1245.
Zhang, J. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by 'death' *Journal of Educational and Behavioral Statistics* **28**, 353–368.
Zhang, J., Rubin, D., and Mealli, F. (2007). Evaluating the effects of job training programs on wages through principal stratification. *Advances in Economics* **21**, 93–118.