Using Statistics to Determine Causal Relationships

Jerome P. Reiter*

1 Introduction

Does a decision to smoke cigarettes increase the likelihood of a person getting lung cancer? Does changing teachers' expectations of a student's performance affect the academic development of that student? Increasingly, such causal questions are being answered with statistics. For both scientists and consumers, it has become important to understand how valid causal studies can be designed and how suspicious studies can be identified. This paper aims to further these understandings by explaining the statistical principles and techniques that underlie valid studies of causal relationships.

The objective of many causal studies–and the objective addressed in this paper–is to measure the effect of some variable on taking one action relative to the effect of taking a different action. For example, how will my headache feel if I take aspirin versus if I do not take aspirin? This differs from the objective of identifying *the* cause of an event (e.g., it was my neighbor's incessantly barking dog that caused my headache). Realistically identifying *the* cause of an event is generally an unattainable goal because of the many variables that affect an outcome: my headache was affected by pressures at work, by the quality of last night's sleep, etc. However, as we shall see, relative causal effects are amenable to statistical analyses (Rubin, 1990).

*Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708. This paper appears in *The American Mathematical Monthly* in January 2000.

2 Formally defining relative causal effects

A statistical framework for determining relative causal effects was constructed by Neyman (1990) and Fisher (1925, 1935) in the context of agricultural studies in which researchers randomly assigned various fertilizers to plots to see how crop yields would respond. Rubin (1974, 1978, 1990) extended the framework to cover settings where researchers do not randomize assignements. Because of the popularity and comprehensibility of the Neyman-Fisher-Rubin framework, our definition of causal effects follows these authors' work. For explanations of other causal frameworks, see Holland (1986), where the Neyman-Fisher-Rubin framework is called the *Rubin Causal Model*, and Cox (1992).

First, we establish some terminology that describes the basics of a causal study. *Treatments* are variables that are conceptually manipulable. For example, in a study addressing ways of reducing people's cholesterol levels, following a vegetarian diet is a treatment because a person's diet can be altered to be vegetarian or non-vegetarian. Conversely, in this study, age is not a treatment because a person's age cannot be manipulated. *Units* are the objects to which the treatments are assigned. In the cholesterol study, the units are the people assigned to follow either the vegetarian or non-vegetarian diet. *Responses* are any variables whose values may have been affected by the treatments, such as cholesterol levels after following a particular diet for six months. *Concomitants* are any variables whose values are unaffected by treatments, such as a unit's age, gender, and cholesterol level before treatment assignment. Putting it all together, a causal study attempts to find the relative effects of the treatments on a response for selected units with given values of concomitants.

We now use the cholesterol study to develop the formal definition of a causal effect. Assume that there is exactly one type of vegetarian diet and one type of non-vegetarian diet. To simplify notation, let's call the vegetarian diet "treatment a" and the non-vegetarian diet "treatment b". For each unit u in the study, there is some time t_1 when the unit begins following one of the diets, and some time t_2 (e.g., six months later) when the cholesterol is measured.

There are two potential outcomes at time t_2 for each unit: the cholesterol level that would be observed

if the unit were exposed to treatment a at time t_1, and the cholesterol level that would be observed if the unit were exposed to treatment b at time t_1. Let's denote these outcomes as Y_{ua} and Y_{ub}, respectively. The only distinction between Y_{ua} and Y_{ub} is exposure to different treatments; so, the only explanation of any difference between Y_{ua} and Y_{ub} is a difference in the effects of the two treatments on cholesterol levels. Thus, if somehow we could simultaneously observe Y_{ua} and Y_{ub}, the quantity $Y_{ua} - Y_{ub}$ would tell us exactly how much the cholesterol level for unit u would change if treatment a were used instead of treatment b. Because of this property, $Y_{ua} - Y_{ub}$ is defined as the causal effect of treatment a relative to treatment b for unit u (Rubin, 1974). When there are n units in the study, one measure of a typical causal effect is the average of these causal effects:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_{ia} - Y_{ib}) = \frac{1}{n}\sum_{i=1}^{n}Y_{ia} - \frac{1}{n}\sum_{i=1}^{n}Y_{ib} = \bar{Y}_a - \bar{Y}_b$$

3 Estimating relative causal effects

In reality, we can assign only one treatment to unit u at time t_1. Thus, we can observe either Y_{ua} or Y_{ub} at time t_1, but not both. As a consequence, we are faced with what Holland (1986) calls the fundamental problem of causal inference: we can never directly observe an individual or average causal effect. However, statistics provides a way around this problem: we can create two groups of units, so that one group receives treatment a and the other group receives treatment b, and then estimate the average causal effect from the observed responses in each group.

A natural estimator of $\bar{Y}_a - \bar{Y}_b$ is the difference in the sample means of each treatment group, $\bar{y}_a - \bar{y}_b$. However, if the groups are not well contructed, $\bar{y}_a - \bar{y}_b$, might estimate the effects of both the treatments and other variables. For example, if the vegetarian group contains more avid exercisers than the non-vegetarian group, and if exercise affects cholesterol levels, then $\bar{y}_a - \bar{y}_b$ estimates the effects of the the different diets and of the different exercise habits. Without strong assumptions about how exersice affects cholesterol levels, we are unable to determine how much of $\bar{y}_a - \bar{y}_b$ is due to the different diets and how much is due to the

different exercise habits.

How should we design the causal study to avoid making such strong and potentially unverifiable assumptions? The advice given by Neyman-Fisher-Rubin is simple and logical: we should construct the treatment groups so that the distributions of all the concomitants that might affect the response are as similar as possible in the two groups. With such balance, the concomitants affect $\bar{y}_a$ and $\bar{y}_b$ by nearly the same amounts, so their effects on $\bar{y}_a - \bar{y}_b$ are negligible. The only remaining explanations of any difference between $\bar{y}_a$ and $\bar{y}_b$ is a difference in the effects of the treatments. In this way, $\bar{y}_a - \bar{y}_b$ in fact estimates the average causal effect of the treatments.

We have now developed the basics of the Neyman-Fisher-Rubin causal framework. The advantage of designing studies within this framework is that, if concomitant information is available prior to applying the treatments, we can check if the study is likely to yield reliable conclusions before ever measuring the responses: we check the balance of the causally-relevant concomitants in the two groups. On the other hand, when causal studies ignore this framework–as in the study described in Example 3.1–their conclusions rely heavily on unverifiable assumptions.

Example 3.1 *A before-and-after study.* Moore and McCabe (1993, pp. 507-508) describe a summer training program designed to improve the speaking skills of teachers of French. The teachers take a French test at the beginning of the summer, attend the summer program, and then take a different French test at the end of the summer. The average of teachers' post-program scores (call this $\bar{y}_{post}$) is significantly higher than the average of their pre-program scores (call this $\bar{y}_{pre}$). It is tempting to attribute this improvement to a causal effect of the program, but there is a flaw in the study's design that undermines any causal conclusions: since every teacher is exposed to the program, there is no way to observe responses at the end of the summer under the no-attendance treatment. Consequently, the validity of $\bar{y}_{post} - \bar{y}_{pre}$ as an estimate of the program's causal effect rests on the assumption that the teachers' final test scores would be similar to their initial test scores had they not attended the program. This assumption could be easily violated. For example, perhaps the teachers were unaccostumed to or nervous about the testing format the first time, and this, rather than

the effect of the program, explains why their scores were lower on the first test. Or, perhaps simply taking the first test motivated the teachers to learn more French independent of the summer program. Even if we believe that taking the first test did not affect scores on the second test, other events may have affected teachers' post-program test scores. For example, perhaps the school system offered incentives to teachers who improved their French, and this rather than the summer program motivated teachers to improve.

The researchers can avoid these issues by assigning some teachers to attend the program and others not to attend, and by giving the initial and final tests to both groups. With this design, any effects due to taking the initial test or the passage of time are present in both groups and are therefore removed from the estimate of the average causal effect.

If the researchers adopt this design, they should balance the causally-relevant concomitants (e.g., years of teaching experience, nationality) in the two groups to isolate the effect of the program. The next two sections describe techniques for creating treatment groups that achieve such balance.

4 The power of randomized experiments

When the researcher controls the assignment of treatments to the units, the study is called an *experiment*. To follow the Neyman-Fisher-Rubin advice, the researcher needs to assign treatments to units (i.e., create treatment groups) in a manner that balances the causally-relevant concomitants. When there are many concomitants, it can be difficult to assign treatments systematically in a way that achieves such balance. Furthermore, even if the researcher manages to balance adequately the concomitants that are observed, the groups might still be unbalanced on unobserved, causally-relevant concomitants.

Amazingly, there is a simple technique that approximately balances both observed and unobserved concomitants in each group: random assignment of treatment to each unit as suggested by Fisher (1935, p. 224). With two treatments, random assignments are determined by tosses of a coin: units whose coin toss is heads are assigned treatment a, and units whose coin toss is tails are assigned treatment b. Constructing treatment groups by this process is basically equivalent to randomly taking two disjoint samples from the

units in the study. Since random samples from the same population tend to have similar characteristics, the two treatment groups should have closely balanced concomitants.

To show this formally, let x_u be the vector of observed and unobserved concomitants of unit u, and let $\mathbf{x}_a = \{x_u : u \text{ assigned treatment } a\}$ represent the collection of concomitants for all units assigned treatment a. Before treatments are assigned, $\mathbf{x}_a$ is a random variable. Its sample space consists of all possible arrangements of the units' concomitants under the study design. For example, when assigning treatment a to two of the four units in a study, $\mathbf{x}_a$ can take on one of six possible outcomes: $(x_1, x_2), (x_1, x_3), (x_1, x_4), (x_2, x_3), (x_2, x_4), (x_3, x_4)$. Because of the randomization, treatment assignment is independent of the concomitants; that is, each unit has the same probability of being assigned treatment a, regardless of the values of its concomitants. Thus, each outcome in the sample space of $\mathbf{x}_a$ is equally likely to occur. By symmetry, the collection of concomitants for all units assigned to treatment b, $\mathbf{x}_b = \{x_u : u \text{ assigned treatment } b\}$, has the same sample space and probability distribution. Hence, since $\mathbf{x}_a$ and $\mathbf{x}_b$ are sampled from the same distribution, the two treatment groups should have closely balanced concomitants.

Example 4.1 *Balance of several concomitants in a real experiment.* The National Supported Work Demonstration was a federally-sponsored study of a job-training program for economically disadvantaged male and female workers (LaLonde, 1986). The study ran in the mid-1970s in ten sites across the United States. Qualified applicants to the program were randomly assigned to one of two groups: 1) a treated group that received the training, and 2) a control group that received no training.

Before randomizing an applicant to a group, the researchers collected background information on the applicant. The means and standard deviations of these concomitants for the 1602 women in the program are shown in Table 1. As is evident from the table, the means and standard deviations of the concomitants are closely balanced. In fact, according to LaLonde (1986), none of the difference between the treated and control group means are statistically significant.

What about the balance of the causally-relevant concomitants not shown in Table 1? Because treatment assignment is independent of all concomitants, there is no reason to think that the randomization acted

Table 1: Concomitants' balance in a randomized experiment

	Treated Group		Control Group	
Variable	Mean	(SD)	Mean	(SD)
Age	33.37	(7.43)	33.63	(7.18)
Years of School	10.30	(1.92)	10.27	(2.00)
Proportion High School Dropouts	.70	(.46)	.69	(.46)
Proportion Married	.02	(.15)	.04	(.20)
Proportion Black	.84	(.37)	.82	(.39)
Proportion Hispanic	.12	(.32)	.13	(.33)
Real Earnings 1 Yr. Before Training ($)	393	(1203)	395	(1149)
Real Earnings 2 Yrs. Before Training ($)	854	(2087)	894	(2240)
Hours Worked 1 Yr. Before Training	90	(251)	92	(253)
Hours Worked 2 Yrs. Before Training	186	(434)	188	(450)
Month of Assignment (Jan. 1978 = 0)	-12.26	(4.30)	-12.30	(4.23)
Number of Observations	800		802	

Note: Data extracted from Table 1 in LaLonde (1986) .

differently on unobserved concomitants. Therefore, although we cannot empirically gauge the sample balance of the unobserved concomitants, we have assurance that their balance is similar to the balance of the observed concomitants.

In addition to closely balancing both observed and unobserved concomitants, random assignment has other related benefits. First, because of randomization, $E(\bar{y}_a - \bar{y}_b) = \bar{Y}_a - \bar{Y}_b$; that is, in expectation the sample average causal effect equals the true average causal effect. This property, which is called *unbiasedness*, implies that $\bar{y}_a - \bar{y}_b$ tends to estimate the right quantity, $\bar{Y}_a - \bar{Y}_b$ (Rubin, 1974). Second, under the null hypothesis that $Y_{ua} - Y_{ub} = k$ for all units u in the study, random assignment induces a pribability distribution on the observable sample average causal effect. Using this distribution, it is possible to test the null hypothesis without making parametric assumptions about the sample average causal effect, e.g., that it follows a normal distribution (Fisher, 1935; Rosenbaum, 1995; Rubin, 1974). Third, when using parametric models to obtain inferences about an average causal effect, randomization simplifies the modeling of causal relationships. With randomized experiments, it is not necessary to model the effect of concomitants on the responses to get valid inferences (Rubin, 1978). Finally, randomization prevents the researcher from purposefully assigning specific units to certain treatments. This helps the researcher avoid cheating that favors one treatment over another

(Rubin, 1978).

Because of these desirable properties, randomized experiments are widely held as the gold standard for studies of causal effects (Cobb and Moore, 1997). For general advice on designing, implementing, and analyzing randomized experiments, see Cox (1958).

5 Estimating causal effects without randomization

Frequently, researchers want to compare the relative effects of several treatments but, for ethical or practical reasons, cannot randomize the treatments to the units. In the statistical literature, such studies are called *observational studies*. For example, a comparison of lung cancer incidence rates for smokers and non-smokers is an observational study, since we cannot assign people to smoke or not to smoke.

The data in an observational study are usually collected from databases that contain units with different treatment exposures, e.g., data on smokers and non-smokers obtained from hospital records. The units in these database populations may have different distributions of causally-relevant concomitants, e.g., the smoking population may be younger than the non-smoking population, and age affects lung cancer rates. Thus, it is desirable to create treatment groups from the database populations that have similar distributions of causally-relevant concomitants rather than simply differencing the averages of smokers' and non-smokers' lung cancer incidence rates from all the units in the database.

Typically, the database contains information on several, but not all, causally-relevant concomitants. One way to balance these observed concomitants is to construct treatment groups from matched pairs, where each half of the pair comes from a different treatment exposure. For example, if for every smoker we include a non-smoker of the same age, gender, and race, the treatment groups will be balanced on these concomitants. As a result, these variables will not affect the comparison of the smoking group's and non-smoking group's lung cancer rates.

When many observed concomitants require balancing, it is often necessary to employ advanced matching techniques, such as matching on propensity scores (Rosenbaum and Rubin, 1983b). In a study with two

treatments a and b, where treatment a is the treatment with fewer exposures in the database, the propensity score is defined as the conditional probability that a unit with observed concomitants $\mathbf{x}_{obs}$ is exposed to treatment a rather than treatment b. Rosenbaum and Rubin (1983b) prove that units with the same propensity score are assigned to treatments independently of $\mathbf{x}_{obs}$; therefore, units with different exposures but identical propensity scores have the same distributions of $\mathbf{x}_{obs}$. These theorems have useful implications for constructing treatment groups: if, for every unit exposed to treatment, a we select a unit exposed to treatment b with nearly the same propoensity score, the treatment groups will be closely balanced on $\mathbf{x}_{obs}$. In real observational studies, the units' true propensity scores are not known and must be estimated from the data. These estimates are usually obtained from a logistic regression of the probability of exposure to treatment a on some function of the observed concomitants. The function of the concomitants is determined by trying various specifications until one is found that produces treatment groups with adequate concomitant balance.

Example 5.1 *Propensity score matching to balance observable concomitants in a real observational study.* Rosenbaum and Rubin (1985) use propensity score matching to create a control group in a study of the effect on children's psychological development of prenatal exposure to barbiturates. The children, born between 1959 and 1961, are culled from a large Danish database. There are 221 children whose mothers took barbiturates and 7,027 children whose mothers did not take barbiturates. The database also contains information on twenty causally-relevant concomitants for each child.

In Table 2, we show the standardized differences in the concomitants' means between the 221 children exposed to barbiturates and potential control groups of children not exposed to barbiturates. The first column shows the standardized differences when the control group contains all 7,027 non-exposed children, and the second column shows the standardized differences when the control group contains all 221 non-exposed children matched on propensity scores. The standardized differences are defined as $(\bar{x}_a - \bar{x}_b)/\sqrt{(s_a^2 + s_0^2)/2}$, where $\bar{x}_a$ and $\bar{x}_b$ are the sample means of the exposed and control groups' concomitants, and s_a^2 and s_0^2 are the sample variances of the 221 exposed and 7,027 non-exposed children's concomitants. The common

Table 2: Improvement in concomitant balance after propensity score matching

Variable	Standardized Differences In Variables' Means ×100 All Controls		Matched Controls
Child Characteristics			
Sex	-7		0
Single/multiple birth (0,1)	-10		-3
Oldest child (yes, no)	-16	*	-5
Child's age at start of study (months)	3		7
Mother Characteristics			
Socioeconomic status (9 ordered categories)	26	**	-10
Mother's education (4 ordered catgories)	15	*	-17
Mother unmarried (yes, no)	-43	**	-7
Mother's age (years)	59	**	-8
Mother's height (5 ordered categories)	18	**	-8
Pregnancy Characteristics			
Weight gain / height^3 (30 values)	0		0
Pregnancy complications (an index)	17	*	-14
Preeclampsia (yes, no)	9		0
Respiratory illness (yes, no)	10		-7
Length of gestation (10 ordered categories)	6		-12
Cigarette consumption, last trimester (10 ordered categories)	-3		0
Other Drugs			
No. exposures to antihistamines (0 - 6)	10		-3
No. exposures to hormones (0 - 6)	28	**	8
Exposed to hormone type 1 (yes, no)	15	*	-2
Exposed to hormone type 2 (yes, no)	19	**	-2
Exposed to hormone type 3 (yes, no)	18	**	-3

Note: Data extracted from Table 1 in Rosenbaum and Rubin (1985).
* indicates $2 \leq |\text{two-sample t-statistic}| < 3$.
** indicates $3 \leq |\text{two-sample t-statistic}|$.

denominator facilitates comparisons of the balance in unmatched and matched control groups.

As we can see from the table, the exposed children are quite different from the 7,027 non-exposed children. For example, children exposed to barbiturates more frequently were the oldest child, were born to unmarried and older mothers, endured complicated pregnancies, and were exposed to other drugs. After propensity score matching, the dramatic differences in many variables are reduced. For example, the standardized differences for the four hormone use variables are reduced from 28, 15, 19, and 18 to a much smaller 8, -2, -2, and -3.

Propensity score matching may not fully remedy concomitant imbalance. For example, after matching on propensity scores in the barbiturate study, there are still moderate differences between exposed and non-exposed groups in mother's education, pregnancy complications, and gestation length. To control the effects of residual imbalances, a frequently-employed approach in observational studies, and in randomized experiments, is to model the relationship between the concomitants and the response. When Y_i is the response and $\mathbf{x}_{obs,i}$ is a $1 \times p$ vector of observed concomitants for unit i, a typical model is the linear regression

$$Y_i = \beta_0 + \beta_1 TREAT_i + \mathbf{x}_{obs,i}\boldsymbol{\beta} + \epsilon_i$$

where $TREAT_i$ equals one if unit i is exposed to treatment a and equals zero if unit i is exposed to treatment b, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and ϵ_i is an error term with a posited distribution (e.g., a normal distribution with mean zero and constant variance). The estimate of β_1 is the estimate of the average causal effect, adjusted for the effects of the concomitants $\mathbf{x}_{obs}$.

Regression adjustments largely remove the effects of residual imbalance when the relationships between the concomitants and the response are accurately modeled. Over a small region of concomitant space, the assumption of linearity between concomitants and response is likely to be reasonable. On the other hand, when the concomitants' distributions are far apart, the linearity assumption is based on unverifiable extrapolations across a wide range of the concomitants. For this reason, regression models are more effective when based on matched groups with similar concomitants than when based on all units in the database (Rubin, 1997).

It is important to note that propensity score matching and regression adjustment do not directly address imbalances in unobserved causally-relevant concomitants. This contrasts with random assignment of treatments, which provides assurance that unobserved background concomitants are not severely imbalanced. In observational studies, it is therefore imperative to test thoroughly the sensitivity of causal conclusions to various specifications of the effects of unobserved concomitants and degrees of imbalance in these concomi-

tants. Such sensitivity tests can be conducted using the techniques in Rosenbaum (1995) and Rosenbaum and Rubin (1983a). Regardless, observational studies are especially vulnerable to the criticism that the estimates of the causal effects are attributable to unobserved concomitants.

Despite their limitations, observational studies are often the only way to address many important causal questions. Thus, we should not remove observational studies from our causal tool box; instead, we should think hard about which concomitants are causally-relevant and do our best to balance them across the treatment groups. For general advice on planning and analyzing observational studies, the reader is referred to the works of Cochran (1983) and Rubin (1984). Another excellent source for information on the analysis of observational studies is Rosenbaum (1995).

6 Other considerations in causal studies

Besides the balance of causally-relevant concomitants, there are other fundamental issues to consider when designing or evaluating causal studies. Paramount is the realism of the study: if a study's conclusions are to have relevance to the real world, the study's conditions must map to the real world. For example, a psychologist might randomly assign subjects to work in an artifically created, stressful situation to study the effects of stress on behavior. This does not necessarily mean that the subjects react similarly in real world stressful situations Moore and McCabe (1993, p. 239). A related concern applies to generalizing the results from the study to broader populations: just because conclusions hold in one population does not mean that they hold in other populations. For example, studies frequently use units selected for convenience, such as volunteers or college students. These units have characteristics that make it risky to extend conclusions to the general population (e.g., volunteers are more cooperative and college students are more educated than the general population).

In addition to these problems, some studies contain hidden biases that undermine their credibility. The researcher expecting one treatment to be more effective might subconsciously, or even consciously, alter the observed responses to favor that treatment. Similarly, the subject who knows she is receiving a new treatment

may be more upbeat than the subject who knows he is receiving a standard treatment, and positive attitude may affect the outcome of interest. To eliminate the potential of these biases to affect the results, treatment assignments are hidden from both researcher and subject in double-blind studies Moore and McCabe (1993, p. 238). Another hidden bias occurs when the treatment for one unit affects the response for other units, thereby distorting estimates of the causal effect. For example, in an agricultural experiment comparing two fertilizers on the same field, the fertilizer from one plot may leach onto another plot. This leaching may affect the plots' yields and, therefore, the conclusions about fertilizer effectiveness. Often studies can be designed to avoid such interference between units: in agricultural experiments, empty strips of land are placed between each plot to minimize the effects of leaching (Cox, 1958).

7 Conclusions

Designing causal studies, and knowing when such studies rely heavily on unstated assumptions, is a tricky undertaking. The framework described in this paper can help in this task. The message of the framework is simple and logical: design studies so that the treatment groups have similar distributions of causally-relevant concomitants. When balanced, the concomitants equally affect both groups' average response and therefore do not affect the estimate of the average causal effect. As we showed, a reliable and easy way of balancing both observable and unobservable concomitants is to assign treatments to units randomly. When this is not possible, matching techniques can be used to create treatment groups with similar distributions of the observable concomitants, but causal conclusions lean on the assumption that such balancing has removed most of the effects of the concomitants from the estimate of the average causal effect. Finally, regardless of whether the causal study is experimental of observational, we must always be on the lookout for unwarranted generalizations and hidden biases that can diminish the study's utility.

References

Cobb, G. W. and Moore, D. S. (1997). Mathematics, teaching, and statistics. *The American Mathematical Monthly* **104**, 801–823.

Cochran, W. G. (1983). *Planning and Analysis of Observational Studies*. John Wiley & Sons, New York.

Cox, D. R. (1958). *Planning of Experiments*. John Wiley & Sons, New York.

Cox, D. R. (1992). Causality: Some statistical aspects. *Journal of the Royal Statistical Society, Series A, General* **155**, 291–301.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, London.

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

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

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review* **76**, 604–620.

Moore, D. S. and McCabe, G. P. (1993). *Introduction to the Practice of Statistics*. W. H. Freeman and Company, New York.

Neyman, J. (1990). On the application of probability theory to agricultural experiments: Essay on statistical principles, Section 9, 1923. Translated in *Statistical Science* **5**, 465–480.

Rosenbaum, P. R. (1995). *Observational Studies*. Springer-Verlag, New York.

Rosenbaum, P. R. and Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B* **45**, 212–218.

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

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* **39**, 33–38.

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

Rubin, D. B. (1978). Bayesian inference for causal effects. *Annals of Statistics* **6**, 34–58.

Rubin, D. B. (1984). William G. Cochran's contributions to the design, analysis, and evaluation of observational studies. In P. S. Rao, ed., *W. G. Cochran's Impact on Statistics*, 37–69. John Wiley & Sons, New York.

Rubin, D. B. (1990). Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference* **25**, 279–292.

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