Pre-Class Response for Lecture 3:

You may have heard of a "Randomized Control Trial" or a "Randomized Experiment." Today, these are often regarded as the "gold standard" of empirical evidence in causality. But this was not always the case, and people today still disagree!

Read up through Section 1 of this paper, which discusses the introduction of randomization into statistics by R.A. Fisher in the early 20th century: *see below*.

Next, read this excerpt from "Instruments, randomization, and learning about development" by Economics Nobel Laureate Angus Deaton: *see below*.

In a short paragraph, summarize the points that Fisher and Deaton are making. Why does Fisher think that randomization is necessary? Why does Deaton think it can be limiting, at least in the context of studying economic development? Which argument are you more sympathetic to a priori? (We will get a lot more into the benefits and drawbacks of experiments in the next few classes).

============

If you are curious, here are the full versions of both articles, but they are not required reading, and may make more sense later in the quarter when we have more background: *see below*.

Journal of the History of Biology (2007) 40:295–325
DOI 10.1007/s10739-006-9119-z

R. A. Fisher and his advocacy of randomization

NANCY S. HALL
University of Delaware Academic Center
Georgetown
DE 19947
USA
E-mail: nhall@udel.edu

Abstract. The requirement of randomization in experimental design was first stated by R. A. Fisher, statistician and geneticist, in 1925 in his book *Statistical Methods for Research Workers*. Earlier designs were systematic and involved the judgment of the experimenter; this led to possible bias and inaccurate interpretation of the data. Fisher's dictum was that randomization eliminates bias and permits a valid test of significance. Randomization in experimenting had been used by Charles Sanders Peirce in 1885 but the practice was not continued. Fisher developed his concepts of randomizing as he considered the mathematics of small samples, in discussions with "Student," William Sealy Gosset. Fisher published extensively. His principles of experimental design were spread worldwide by the many "voluntary workers" who came from other institutions to Rothamsted Agricultural Station in England to learn Fisher's methods.

Keywords: Charles Darwin, Charles S. Peirce, experimental design, probable error, Ronald A. Fisher, randomization, Rothamsted, small samples, "student", William S. Gosset

Introduction

Ronald A. Fisher (1890–1962) originated many of the statistical techniques in use today. Statisticians will recognize analysis of variance; the F ratio is named in his honor. A dramatic change in the design of experiments occurred in the first half of the twentieth century, from systematic design to the randomized design of experiments. Fisher was largely responsible for these changes in experimentation and statistics. While some randomization occurred in experimental work prior to Fisher's, Fisher was responsible for a major change in the way randomization was, and still is, used, in many areas of research. He advocated randomization even in experiments that could be conducted otherwise; that is, he argued against using some *non-random systematic* plan chosen by the experimenter.

Fisher's *reasons* for advocating randomization are clear. In many of his publications, and in his correspondence, he was explicit about the need to randomize and the peril of not doing so. But his own *development* of the concept, how he came to recognize that randomization is a requirement of good experimental design, has been shrouded. Nowhere in his published work does such a discussion appear, and there are no direct statements in his surviving correspondence. However, a reading of Fisher's correspondence, especially that with "Student" (William Gosset), in combination with a reading of Fisher's early publications, suggests a circumstantial but strong pattern. This paper will suggest that Fisher saw what was achieved by the randomness of sampling analyses, and that he imported randomness from sampling into experimental design; randomness that is a *property* of sampling became a *requirement* of experimental design.

Part 1 will define randomization and briefly consider its advantages. A biographical note on Fisher will supply historical context. Part 2 will summarize experimental design prior to Fisher, especially in agriculture. Part 3 will consider some early instances of randomizing, including the nineteenth century work of Charles Sanders Peirce (Peirce and Jastrow, 1885) in response to that of Gustav Fechner (1860). The spread of interest in psychophysics, first in England and then in the United States, raised questions that led to experiments concerning psychical phenomena. Part 3 will also examine two recent interpretations of the origins of randomization, those of Ian Hacking (1988) and Trudy Dehue (1997). Part 4 will consider what led Fisher to advocate randomization, as indicated in his correspondence and his published work: his reliance on *n*-dimensional geometry and his continuing interest in the mathematical analysis of small samples. Then will come the chronology of his promotion of randomization, and the spread of Fisherian experimental techniques around the world. This dispersion was due not only to Fisher's publications but to the training of the many "voluntary workers," funded by their home institutions, who came to work and study with Fisher. Those identified so far, and their home affiliations, are listed in the Appendix.

Part 1. Fisher and the Importance of Randomization

During the design of an experiment, randomization is a consideration when an individual subject (or plot of ground) is being assigned to a group and also when a group is being assigned to a treatment.

> A method or procedure is a random method of selection just in case (i) it is possible for each of the n assignments to be selected by the use of that method and (ii) the statistical probability or chance of any one of the n assignments being selected on an application of that method is the same value 1/n as the statistical probability of any other assignment being selected. (Levi, 1983, p. 449)

Fisher applied such a definition in *The Design of Experiments* (1935); he was explicit about assigning treatments in field experiments in agriculture.

> ...and this assignment is made at random. This does not mean that the experimenter writes down the names of the varieties...in any order that may occur to him, but that he carries out a physical experimental process of randomisation, using means which shall ensure that each variety has an equal chance of being tested on any particular plot of ground. A satisfactory method is to use a pack of cards numbered from 1 to 100....(Fisher, 1935, p. 56)

The randomization had to be a "physical experimental process," whether by using a deck of cards or a table of random numbers; the experimenter might just flip a coin, if there were only two choices of assignment.

In an early chapter of *The Design of Experiments*, Fisher laid the groundwork for the randomization requirement by pointing out that, no matter how diligently the experimenter works to control the research conditions, "it would be impossible to present an exhaustive list of such possible differences appropriate to any one kind of experiment, because the uncontrolled causes which may influence the result are always strictly innumerable." (Fisher, 1935, p. 21) The "essential safeguard" is a "full procedure of randomization, by which the validity of the test of significance may be guaranteed against corruption by the causes of disturbance which have not been eliminated." (p. 23) In other words, it is not possible for the experimenter to eliminate all sources of variation other than the one under test; there will invariably be slight differences, such as temperature, soil fertility, drainage, or even time of treatment between those treated first and those treated last. Randomization is a means of eliminating bias in the results due to uncontrolled differences in the experimental conditions. According to Fisher, "Randomization properly carried out...relieves the experimenter from the anxiety of considering and estimating the magnitude of the innumerable causes by which his data may be disturbed." (p. 49)

In an ideal world, agricultural field experiments would be performed on plots known to be identical in all respects. Such ideal conditions can almost be achieved in laboratories in some other research disciplines; it is possible to control the medium in the test tubes or petri dishes. In such cases, there is no need for an estimate of error; the error due to uncontrolled variation in the environment is known to be zero (or vanishingly small). Then all observed variation can be ascribed to the experiment itself. But such control of conditions is impossible in agricultural fields, as well as in many other areas of research. Consequently two problems immediately arise: we need to keep error at a minimum *and* we need to be able to estimate that error. In Fisher's words, we have "the two *desiderata* of the *reduction of error* and of the *valid estimation* of error...." (Fisher, 1926, p. 508, his emphasis)

Researchers had two main objections to randomizing: the randomizing procedure was, first, unnecessary and, second, inconvenient. Many viewed randomization as unnecessary because experiments, especially in agriculture, had been carried out for generations using a systematic approach and the results were regarded as reliable. A test of significance was seen as unnecessary if the observed difference between treatments was dramatic. Yet, if pressed, the researchers using a systematic layout would admit, and indicated in their published accounts, that part of the intent of the layout was to eliminate bias in the fertility gradient, those variations of drainage, soil type, etc., that they could see in the experimental fields. Most systematic arrangements attempted to achieve a balance of sorts. Many articles in both the *Journal of Agricultural Science* and the *Journal of the Ministry of Agriculture* from 1860 to about 1925, describing field experiments, devoted several paragraphs to the efforts that were made to eliminate unwanted sources of variation. Inconvenience was the second objection to randomization, inconvenience both in the initial treatment and/or planting, and in the harvesting. The view was that in both phases a randomized experiment would take more time than a systematic method and also that mistakes would be much more likely – human error in either application.

Fisher first came on the agricultural scene in 1919. The most comprehensive source of biographical information is the book written by his daughter, Joan Fisher Box; *R. A. Fisher: The Life of a Scientist*, published in 1978. Fisher was born in February 1890, in London, England. At Gonville and Caius College at Cambridge he studied mathematics; he then spent a post-graduate year studying quantum theory and statistical mechanics under James Jeans and the theory of errors (now called the normal distribution or the bell shaped curve) under F. J. M.

Stratton. (Tankard, 1984) Ineligible for military service due to his poor vision, during World War I Fisher became a teacher of mathematics and physics at several boys' public schools. At the end of the war, he took a six-month position at Rothamsted Agricultural Station, in Harpenden, about 25 miles northeast of London. He stayed for 14 years.

Fisher's first advocacy in print of randomization came in 1925, but his correspondence makes it clear that he had come to this conclusion by the summer of 1924. His development of randomization before 1924 and his work on randomization from 1924 to 1935 will be of most concern here.

Fisher died in Australia in 1962; his papers, including sixteen feet of correspondence, are archived at the Barr Smith Library at the University of Adelaide.In the archives is a five volume set of letters from W. S. Gosset (Student) to Fisher, 193 letters in all. Student and Fisher had a long and close personal relationship. They first corresponded in 1915 and first met personally in 1918. (Gosset was employed as a brewer by Guinness in Ireland. Guinness would not allow its employees to publish under their own names; Gosset adopted the pseudonym Student. As a result, those entering the study of statistics are somewhat mystified at the nomenclature; one of the most important statistical tools they encounter is known everywhere as Student's *T*-Test.) It seems that Student did not keep Fisher's letters; diligent searching has turned up only a few. Fisher kept Student's letters, and in 1937, when Student died and Fisher was asked to write a memorial tribute, apparently Fisher then organized the letters, numbered them, and wrote a brief commentary on each. Four volumes are letters and the fifth is Fisher's commentaries. The "give and take" in these letters is fascinating; it is also useful in our exploration of Fisher's development of randomization. It is clear that by the summer of 1924, Student was reading the proofs of Fisher's *Statistical Methods for Research Workers* and liking most of what he read.

Rothamsted Agricultural Station is in Harpenden, England, about 25 miles north of London. The statistical library there houses Fisher's own 'Millionaire' mechanical calculating machine,[1] purchased for him (at then great expense) by the station when he began work there in 1919. The main library has complete sets of the *Rothamsted Experimental Station Report*, the *Rothamsted Memoirs on Agricultural Science*, and *Records of the Rothamsted Staff*.

[1] The calculating machine is about three feet long and a foot wide; it resembles an electronic piano but with numbered buttons and round numbered dials instead of keys.

In section 4 of this paper, I shall argue that, *under ideal circumstances*, randomized evaluations of projects are useful for obtaining a convincing estimate of the average effect of a program or project. The price for this success is a focus that is too narrow and too local to tell us "what works" in development, to design policy, or to advance scientific knowledge about development processes. Project evaluations, whether using randomized controlled trials or nonexperimental methods, are unlikely to disclose the secrets of development nor, unless they are guided by theory that is itself open to revision, are they likely to be the basis for a cumulative research program that might lead to a better understanding of development. This argument applies a fortiori to instrumental variables strategies that are aimed at generating quasi-experiments; the value of econometric methods cannot and should not be assessed by how closely they approximate randomized controlled trials. Following Nancy Cartwright (2007a, 2007b), I argue that evidence from randomized controlled trials can have no special priority. Randomization is not a gold standard because "there is no gold standard" Cartwright (2007a.) Randomized controlled trials cannot automatically trump other evidence, they do not occupy any special place in some hierarchy of evidence, nor does it make sense to refer to them as "hard" while other methods are "soft." These rhetorical devices are just that; metaphor is not argument, nor does endless repetition make it so.

More positively, I shall argue that the analysis of projects needs to be refocused toward the investigation of potentially generalizable mechanisms that explain why and in what contexts projects can be expected to work. The best of the experimental work in development economics already does so because its practitioners are too talented to be bound by their own methodological prescriptions. Yet there would be much to be said for doing so more openly. I concur with Ray Pawson and Nick Tilley (1997), who argue that thirty years of project evaluation in sociology, education, and criminology was largely unsuccessful because it focused on *whether* projects worked instead of on *why* they worked.

Journal of the History of Biology (2007) 40:295–325
DOI 10.1007/s10739-006-9119-z

R. A. Fisher and his advocacy of randomization

NANCY S. HALL
University of Delaware Academic Center
Georgetown
DE 19947
USA
E-mail: nhall@udel.edu

Abstract. The requirement of randomization in experimental design was first stated by R. A. Fisher, statistician and geneticist, in 1925 in his book *Statistical Methods for Research Workers.* Earlier designs were systematic and involved the judgment of the experimenter; this led to possible bias and inaccurate interpretation of the data. Fisher's dictum was that randomization eliminates bias and permits a valid test of significance. Randomization in experimenting had been used by Charles Sanders Peirce in 1885 but the practice was not continued. Fisher developed his concepts of randomizing as he considered the mathematics of small samples, in discussions with "Student," William Sealy Gosset. Fisher published extensively. His principles of experimental design were spread worldwide by the many "voluntary workers" who came from other institutions to Rothamsted Agricultural Station in England to learn Fisher's methods.

Keywords: Charles Darwin, Charles S. Peirce, experimental design, probable error, Ronald A. Fisher, randomization, Rothamsted, small samples, "student", William S. Gosset

Introduction

Ronald A. Fisher (1890–1962) originated many of the statistical techniques in use today. Statisticians will recognize analysis of variance; the F ratio is named in his honor. A dramatic change in the design of experiments occurred in the first half of the twentieth century, from systematic design to the randomized design of experiments. Fisher was largely responsible for these changes in experimentation and statistics. While some randomization occurred in experimental work prior to Fisher's, Fisher was responsible for a major change in the way randomization was, and still is, used, in many areas of research. He advocated randomization even in experiments that could be conducted otherwise; that is, he argued against using some *non-random systematic* plan chosen by the experimenter.

Fisher's *reasons* for advocating randomization are clear. In many of his publications, and in his correspondence, he was explicit about the need to randomize and the peril of not doing so. But his own *development* of the concept, how he came to recognize that randomization is a requirement of good experimental design, has been shrouded. Nowhere in his published work does such a discussion appear, and there are no direct statements in his surviving correspondence. However, a reading of Fisher's correspondence, especially that with "Student" (William Gosset), in combination with a reading of Fisher's early publications, suggests a circumstantial but strong pattern. This paper will suggest that Fisher saw what was achieved by the randomness of sampling analyses, and that he imported randomness from sampling into experimental design; randomness that is a *property* of sampling became a *requirement* of experimental design.

Part 1 will define randomization and briefly consider its advantages. A biographical note on Fisher will supply historical context. Part 2 will summarize experimental design prior to Fisher, especially in agriculture. Part 3 will consider some early instances of randomizing, including the nineteenth century work of Charles Sanders Peirce (Peirce and Jastrow, 1885) in response to that of Gustav Fechner (1860). The spread of interest in psychophysics, first in England and then in the United States, raised questions that led to experiments concerning psychical phenomena. Part 3 will also examine two recent interpretations of the origins of randomization, those of Ian Hacking (1988) and Trudy Dehue (1997). Part 4 will consider what led Fisher to advocate randomization, as indicated in his correspondence and his published work: his reliance on *n*-dimensional geometry and his continuing interest in the mathematical analysis of small samples. Then will come the chronology of his promotion of randomization, and the spread of Fisherian experimental techniques around the world. This dispersion was due not only to Fisher's publications but to the training of the many "voluntary workers," funded by their home institutions, who came to work and study with Fisher. Those identified so far, and their home affiliations, are listed in the Appendix.

Part 1. Fisher and the Importance of Randomization

During the design of an experiment, randomization is a consideration when an individual subject (or plot of ground) is being assigned to a group and also when a group is being assigned to a treatment.

> A method or procedure is a random method of selection just in case (i) it is possible for each of the n assignments to be selected by the use of that method and (ii) the statistical probability or chance of any one of the n assignments being selected on an application of that method is the same value 1/n as the statistical probability of any other assignment being selected. (Levi, 1983, p. 449)

Fisher applied such a definition in *The Design of Experiments* (1935); he was explicit about assigning treatments in field experiments in agriculture.

> ...and this assignment is made at random. This does not mean that the experimenter writes down the names of the varieties...in any order that may occur to him, but that he carries out a physical experimental process of randomisation, using means which shall ensure that each variety has an equal chance of being tested on any particular plot of ground. A satisfactory method is to use a pack of cards numbered from 1 to 100....(Fisher, 1935, p. 56)

The randomization had to be a "physical experimental process," whether by using a deck of cards or a table of random numbers; the experimenter might just flip a coin, if there were only two choices of assignment.

In an early chapter of *The Design of Experiments*, Fisher laid the groundwork for the randomization requirement by pointing out that, no matter how diligently the experimenter works to control the research conditions, "it would be impossible to present an exhaustive list of such possible differences appropriate to any one kind of experiment, because the uncontrolled causes which may influence the result are always strictly innumerable." (Fisher, 1935, p. 21) The "essential safeguard" is a "full procedure of randomization, by which the validity of the test of significance may be guaranteed against corruption by the causes of disturbance which have not been eliminated." (p. 23) In other words, it is not possible for the experimenter to eliminate all sources of variation other than the one under test; there will invariably be slight differences, such as temperature, soil fertility, drainage, or even time of treatment between those treated first and those treated last. Randomization is a means of eliminating bias in the results due to uncontrolled differences in the experimental conditions. According to Fisher, "Randomization properly carried out...relieves the experimenter from the anxiety of considering and estimating the magnitude of the innumerable causes by which his data may be disturbed." (p. 49)

In an ideal world, agricultural field experiments would be performed on plots known to be identical in all respects. Such ideal conditions can almost be achieved in laboratories in some other research disciplines; it is possible to control the medium in the test tubes or petri dishes. In such cases, there is no need for an estimate of error; the error due to uncontrolled variation in the environment is known to be zero (or vanishingly small). Then all observed variation can be ascribed to the experiment itself. But such control of conditions is impossible in agricultural fields, as well as in many other areas of research. Consequently two problems immediately arise: we need to keep error at a minimum *and* we need to be able to estimate that error. In Fisher's words, we have "the two *desiderata* of the *reduction of error* and of the *valid estimation* of error...." (Fisher, 1926, p. 508, his emphasis)

Researchers had two main objections to randomizing: the randomizing procedure was, first, unnecessary and, second, inconvenient. Many viewed randomization as unnecessary because experiments, especially in agriculture, had been carried out for generations using a systematic approach and the results were regarded as reliable. A test of significance was seen as unnecessary if the observed difference between treatments was dramatic. Yet, if pressed, the researchers using a systematic layout would admit, and indicated in their published accounts, that part of the intent of the layout was to eliminate bias in the fertility gradient, those variations of drainage, soil type, etc., that they could see in the experimental fields. Most systematic arrangements attempted to achieve a balance of sorts. Many articles in both the *Journal of Agricultural Science* and the *Journal of the Ministry of Agriculture* from 1860 to about 1925, describing field experiments, devoted several paragraphs to the efforts that were made to eliminate unwanted sources of variation. Inconvenience was the second objection to randomization, inconvenience both in the initial treatment and/or planting, and in the harvesting. The view was that in both phases a randomized experiment would take more time than a systematic method and also that mistakes would be much more likely – human error in either application.

Fisher first came on the agricultural scene in 1919. The most comprehensive source of biographical information is the book written by his daughter, Joan Fisher Box; *R. A. Fisher: The Life of a Scientist*, published in 1978. Fisher was born in February 1890, in London, England. At Gonville and Caius College at Cambridge he studied mathematics; he then spent a post-graduate year studying quantum theory and statistical mechanics under James Jeans and the theory of errors (now called the normal distribution or the bell shaped curve) under F. J. M.

Stratton. (Tankard, 1984) Ineligible for military service due to his poor vision, during World War I Fisher became a teacher of mathematics and physics at several boys' public schools. At the end of the war, he took a six-month position at Rothamsted Agricultural Station, in Harpenden, about 25 miles northeast of London. He stayed for 14 years.

Fisher's first advocacy in print of randomization came in 1925, but his correspondence makes it clear that he had come to this conclusion by the summer of 1924. His development of randomization before 1924 and his work on randomization from 1924 to 1935 will be of most concern here.

Fisher died in Australia in 1962; his papers, including sixteen feet of correspondence, are archived at the Barr Smith Library at the University of Adelaide.In the archives is a five volume set of letters from W. S. Gosset (Student) to Fisher, 193 letters in all. Student and Fisher had a long and close personal relationship. They first corresponded in 1915 and first met personally in 1918. (Gosset was employed as a brewer by Guinness in Ireland. Guinness would not allow its employees to publish under their own names; Gosset adopted the pseudonym Student. As a result, those entering the study of statistics are somewhat mystified at the nomenclature; one of the most important statistical tools they encounter is known everywhere as Student's *T*-Test.) It seems that Student did not keep Fisher's letters; diligent searching has turned up only a few. Fisher kept Student's letters, and in 1937, when Student died and Fisher was asked to write a memorial tribute, apparently Fisher then organized the letters, numbered them, and wrote a brief commentary on each. Four volumes are letters and the fifth is Fisher's commentaries. The "give and take" in these letters is fascinating; it is also useful in our exploration of Fisher's development of randomization. It is clear that by the summer of 1924, Student was reading the proofs of Fisher's *Statistical Methods for Research Workers* and liking most of what he read.

Rothamsted Agricultural Station is in Harpenden, England, about 25 miles north of London. The statistical library there houses Fisher's own 'Millionaire' mechanical calculating machine,[1] purchased for him (at then great expense) by the station when he began work there in 1919. The main library has complete sets of the *Rothamsted Experimental Station Report*, the *Rothamsted Memoirs on Agricultural Science*, and *Records of the Rothamsted Staff*.

[1] The calculating machine is about three feet long and a foot wide; it resembles an electronic piano but with numbered buttons and round numbered dials instead of keys.

Now we will step back in time, to agriculture in nineteenth century England and experimentation before the era of Fisher, describing systematic design and its problems.

Part 2. Experimental Design Prior to Fisher

The technical books and journals on soils, fertilizers and field experiments from the nineteenth and early twentieth centuries are of interest from a statistical or design point of view in two ways. First, careful reading of any one may provide a snapshot of statistical practices, methods of reporting results and that particular author's attitude toward the question of randomization in experiments. Second, by comparing several it is possible to establish a chronology of change, in practices and in attitude.

Agricultural field experiments on a large scale were under way by the middle of the nineteenth century, at first financed privately and later, beginning in 1910, supported by the British government. The methodology governing experimental design prior to Fisher's exposition of randomization faced the difficulty of variation: if two plots were fertilized at different intensities and one did better than the other, was that difference due to the fertilizer or would one plot have been superior anyhow? The usual practice, to avert this, was to select the two plots to be as similar as possible and then make the assumption that they were identical for purposes of the experiment. Then two more difficulties became apparent. A value judgment was now involved; the researcher had to decide if the two plots were sufficiently alike, but the drainage might be slightly different, or the light and shadow, or even the wind.

By the end of the nineteenth century, agriculturalists had come to realize that one useful strategy was that of replication. However, repetitions immediately gave rise to another consideration: how to interpret the data when, as usually happened, plots treated exactly alike did not produce exactly the same results.

The probable error was the distance from the mean that encompassed 1/4 or 25% of the data on each side of the mean, 50% total, and the question of managing the experimental error was much on the minds of researchers early in the twentieth century.

> A very cursory examination of the results of any set of field trials will serve to show that a pair of plots similarly treated may be expected to yield considerably different results, even when the soil

> appears to be uniform and the conditions under which the experiment is conducted are carefully designed to reduce errors in weighing and measurement ...[T]he *probable error* attaching to a single plot is in the neighborhood of plus or minus 10 per cent.... [T]he chances are even against the result being within 10 per cent of the truth,... the average result obtained by a number of experiments... (Mercer and Hall, 1911, p. 107, my emphasis)

In other words, uniformity trials (experiments where plots were treated identically) had shown that the data from any single plot, compared with the average over the field, might be within 10% of the average or might be more than 10% of the average, and "more than" was just as likely as "within." Fisher's methods, including randomization, were to make possible a much smaller experimental error.

According to Russell (1966), concern about experimental error in agricultural experimentation did not arise until after 1906, when S. H. Collins at Armstrong College published research on mangolds, a kind of beet. Collins expressed his results mathematically, adding together factors from crop, variety, soil, and so on. This provoked A. D. Hall of Rothamsted to protest that an estimate of error was needed; this had not previously been attempted in agricultural research.

At Cambridge T. H. Wood was acquainted with a noted astronomer, F. J. M. Stratton (with whom Fisher later studied). In what may be considered a landmark paper in English agriculture, Wood and Stratton (1910), in "The Interpretation of Experimental Results," applied the astronomer's techniques of estimating error to several sets of agricultural data. (Russell, 1966, pp. 205–206) In the section on probable error the authors themselves noted the oddity of the situation.

> It might seem at first that no two branches of study could be more widely separated than Agriculture and Astronomy....The astronomer's measurements come up short of absolute accuracy because of a great number of atmospheric conditions, each of which is equally likely to make any one result high or low. He has to obviate this unavoidable lack of accuracy by making many independent observations and taking their average. This is, or should be, the method followed by the agriculturalist....The astronomer, being a mathematician, has devised a method of estimating the accuracy of his averages, which he invariably applies with great advantage. (Wood and Stratton, p. 425)

They then applied the astronomers' technique of least squares and probable error, using as data the percentage of dry matter in each of 10 mangolds. The average of the 10 was 14.3 and calculations produced a probable error of ± 1.3. The explanation was given that "half of the results should differ from the mean by less than the probable error, the other half by more." (Wood and Stratton, p. 426)

Here things stood in the estimation of error, until Fisher's statistical techniques introduced in the 1920's had the effect of consigning the concept of probable error to the dustbin of mathematical history. Fisher recognized early the connection that could be considered between variability in a sample and the errors of observation that are routinely considered in astronomy.

> If a measurement...is made upon a sufficiently large sample...the measurements are usually found to be grouped symmetrically about a mean value, the average...of the sample. The deviations from this average follow very closely...the law of errors....[T]he frequencies with which deviations of different magnitudes occur are related in the same way as the frequencies of errors of observation. Consequently the amount of variability may be measured, as errors of observation are habitually measured...this mean square deviation being strictly comparable to the mean square error used in astronomy and in all physical science. (Fisher, 1918, p. 213)

Now we return to the problem of attempting to minimize the error resulting from differences in the experimental material, in this case the field plots. A theoretical solution to the difference between plots would be to use the same plot twice – raise a crop using one level of fertilizer, harvest, and then replant the same crop using the second level of fertilizer. But now the difference would be one of time rather than location. No two growing seasons are identical as to temperature and rainfall; once again the researcher must fall back on assumptions, this time as to the relative effects of the differing growing seasons.

Systematic experimental designs were the accepted practice until Fisher's work became known. Minimizing the probable error (or, the same goal, minimizing the standard deviation) was an essential part of Fisher's work as he developed and refined both statistical methods and randomized experimental designs, during his years at Rothamsted.

One of the features for which Rothamsted is well known began in 1852, the long-term wheat experiments – the planting of wheat of the same variety with the same manurial treatment year after year on the same plot. Similar plantings were later started for barley, mangolds, clover and grass for hay. By the time Fisher arrived at Rothamsted in 1919, some of these experimental plots had been maintained – manured, planted, harvested, yield measured and recorded – for more than 60 years. Also, the records of temperature and rainfall had been maintained during all that time.

E. John Russell became director at Rothamsted in 1912. Russell, after some time, set about to introduce the then new subject of statistics. "One of the first things I had done when I became Director was to look in the cupboards, where I found great masses of data from the field experiments. The figures were good, all honestly and carefully taken and recorded. But they had never been worked up and I knew that I was incompetent to undertake the task." (Russell, 1956, p. 131) So he hired Ronald A. Fisher, age 29, to set up a statistics department.

Fisher was not the first to randomize in experimentation. In the next section will be earlier uses of randomization, effective experimental designs that produced practical results, but from which nothing more was heard. The seed of the idea of randomizing was planted, and in some cases flourished briefly, but then died out.

Part 3. Emergence of Randomization Before Fisher

Randomization appeared in psychopsychics research in the late nineteenth century and continued actively in England for many years. Historians accord the distinction of being the first to randomize in experimentation to Charles Sanders Peirce, who reported his experiments in the 1885 *Memoirs of the National Academy of Sciences.* He described his experimental design in detail, including randomization, but nothing came of it; no one else followed.

At Johns Hopkins University in the winter of 1883–1884 Charles Sanders Peirce and a student of his, Joseph Jastrow, carried out experiments to test a human subject's ability to detect slight differences in weight. Gustav Fechner, in *Elemente der Psychophysik* (1860), had maintained that a lower limit of perception exists, below which a subject is not able to distinguish a slight difference between two sensations of weight. Peirce questioned the assumption that this "Unterschiedsschwelle" exists and structured experiments to test the assumption. He described his experimental methods and his results in "Small Differences of Sensation" for the *Memoirs of the National Academy of Sciences* in 1885.

> If there be a least perceptible difference, then when two excitations differing by less than this are presented to us, and we are asked to judge which is the greater, we ought to answer wrong as often as right in the long run. Whereas, if [not]...we...ought to answer right oftener than wrong....(Peirce and Jastrow, 1885, p. 73)

Peirce created an elaborate mechanical apparatus; a subject first experienced a weight of one kilogram and then a second weight either slightly heavier or slightly lighter than the first. The subject was required to offer an opinion; that is, he was not allowed to say that he could not tell. Also, having declared an opinion, he then had to rate his own degree of confidence in his opinion, from 0 to 3. Of interest to us is Peirce's use of randomization to determine whether weight should be added or subtracted from the one kilogram: he and Jastrow used ordinary red and black playing cards; the operator either increased or decreased the weight according to the color of the next card. Peirce and Jastrow alternated as subject and operator; with 150 trials at a sitting they accumulated data from almost 5000 trials. In the approximately 3000 trials where the subject declared he had zero confidence in his opinion – he believed he could not tell any difference and was just guessing – that guess was correct about 3 out of 5 times; Peirce therefore concluded that the "Unterschiedsschwelle" does not exist.

Peirce included a detailed description of the use of the decks of cards for randomization, but he seems not to have recognized that his randomization was in any way special or unique in experimental practice. In "Small Differences in Sensation" he described the use of the cards, the mechanical apparatus, the methods and timing of data collection and the mathematical analysis. Randomization was not worthy of mention in his concluding discussion. Furthermore, randomization was also not mentioned in the issue of *Science* that reported on papers presented at the fall 1884 meeting of the National Academy of Sciences. Writing in 1978, Stephen M. Stigler remarked, "The Peirce-Jastrow experiment is the first of which I am aware where the experimentation was performed according to a precise, mathematically sound randomization scheme!" (Stigler, 1978, p. 248)

In "Telepathy and Randomization," Ian Hacking (1988) considered Peirce's experiments to be the first use of "artificial randomization," artificial in the sense that a device, a deck of playing cards, was used to determine the next trial, rather than the whim of the experimenter. Hacking also described a number of instances of "faltering use" of randomization in psychophysics (today: experimental psychology), telepathy, and psychical research in the 1880's. But Hacking credited

Fisher with the promulgation of the randomization principle, and I will argue that the evidence supports Hacking's assessment. According to Hacking, Fisher knew of the psychical research activities, including randomizing, but was not influenced by them. Fisher's requirement of randomization in experimental design was part of "his own vision of the conceptual foundations of statistics." (Hacking, 1988, p. 449)

The evidence supports Hacking's conclusions, but not those of Trudy Dehue; for her the psychological community rather than Fisher was responsible for the emergence of randomization and its widespread adoption. Dehue (1997), in "Deception, Efficiency and Random Groups," argued that Fisher's random grouping was in fact being practiced earlier in the twentieth century in experiments in psychology. The subtitle of her article was "Psychology and the Gradual Origin of the Random Group Design." Her position was that randomization in the design of experiments could not be attributed to any one person or era. Dehue claimed that random group design came about gradually within the experimental psychology community. In her opinion, neither psychophysics nor psychology played any role in Fisher's development of his statistical methods, including randomization. In fact, the effect went the other way: she documented the rapid adoption of Fisher's methods (which included randomizing) by the experimental psychology community, whose projects had become so involved that the results were difficult to handle.

Using essentially the same sources as Hacking (1988) with the addition of some materials from education, Dehue reached a conclusion very different from his:

> I argue... that the random group design was advanced in psychology before Fisher introduced it in agriculture and that in this context it was the unplanned outcome of a lengthy historical process rather than the instantaneous creation[2] of a single genius. (Dehue, 1997, p. 655)

Her evidence, after citing Hacking (1988) extensively for the experiments in telepathy mentioned above, was only that psychologists found it expedient to use matched *groups* of subjects when the then-customary matched *pairing of individuals* was inconvenient or too expensive. Traditionally, pairs of subjects were formed that were "matched" on "each possible contaminating factor." (The twin assumptions here were that, first, this could be achieved, and, second, that then the observed

[2] I am not aware that anyone else has claimed that Fisher's randomization requirement was an "instantaneous creation."

differences in experimental results could be attributed to the property being considered.) Dehue cited McCall's 1923 *How to Experiment in Education*, in which he advocated a less costly alternative to matched pairs: have a large number of experimental subjects, students appropriate for the trial, and separate them into two groups by some randomizing means; one group then served as the control and the other participated in the experiment. I suggest that this one instance, McCall's suggestion (in 1923, 1 year before Fisher's writing his first book advocating randomization) does not support her claim of the randomized group design being the outcome of a lengthy historical process. Also, experimentation in both psychology and education became more elaborate in the early decades of this century; there was an increased concern for managing uncontrolled variables and also worry about whether the observed effect was the result of the experimenter's efforts or of something else. This increasing complexity might support her claim of "gradual emergence" of random group design but not its spread to other disciplines. My thesis is that Fisher was responsible for the lasting introduction and later widespread acceptance of randomization in experimental design.

Part 4. What Led Fisher to Randomization

According to Fisher's daughter, "It is uncertain just when Fisher made the intuitive leap by which he recognized the principle of randomization." (Box, 1978, p. 147) She surmised that it must have been between 1918 and 1923; this suggested the desirability of making a careful search of Fisher's correspondence and published papers in those years.

Fisher had begun consideration of the statistics of small samples by 1915, the beginning of his correspondence with Student. This involvement with small samples is reflected in a number of Fisher's published papers through 1923. By the summer of 1924 Student was reading the proofs of *Statistical Methods for Research Workers*, and objecting to Fisher's requirement of randomization in experimental design; in his view, the advantages of randomizing were outweighed by the increased complexity and thus the increased possibility of making mistakes in actually carrying out the experiment and collecting the data.

The chronology of Fisher's advocacy of randomization in experimental design can be traced by examining the interplay of published articles by Fisher and his contemporaries – Student (William Sealy Gosset), Karl Pearson, E. John Russell, and Frank Yates, among others – and the voluminous correspondence between them. Fisher enjoyed vigorous interaction with many.

Fisher was guided by two principles. First, one must make a distinction between the statistic (a summary of the data) and its reliability. Second, one must have a measure of this reliability; the procedure by which the sample was gathered must allow for this measure. In other words, having sampled the population (of the acidity in vats of beer or heights of soldiers – it does not matter what) and having arrived at a number that is an estimate of the property of interest, one must immediately ask how likely it is that this is a good portrait of the population; i.e., how likely it is that more samples of the same size would lead to the same portrait. Fisher called this reliability number the "test of significance."

Now we will examine Fisher's visualization of *n*-dimensional geometry to generate the mathematics that he then used in several different papers. In some cases the project was to validate the work of others; in many cases he utilized geometry to extend the work of others and he then went on to arrive at something new. We will be particularly interested in the connection between random sampling and geometry.

In 1915, in *Biometrika*, Fisher published a very detailed mathematical discussion of Student's statistical work on small samples. In 1908 Student had published two papers, one on the probable error of the mean of a small sample, and one on the probable error of the coefficient of correlation of two small samples (Student, 1908a, b). Student, in the Guinness brewery, needed an estimation of the accuracy of the mean of a small sample – perhaps numbering as small as two or three or four. In his paper (Fisher, 1915), Fisher said essentially that Student's analysis was right. Fisher's initial appeal was geometric.

> This result, although arrived at by empirical methods, was established almost beyond reasonable doubt...[but] the form establishes itself *instantly* when the distribution of the sample is viewed geometrically. (Fisher, 1915, p. 507, emphasis added)

What followed were two pages of geometry, discussed in *n* dimensions but, for the reader's benefit pictured in three dimensions, and then *twelve pages* of integral calculus; so much for "*instantly*."[3]

[3] It is not an original observation to note that Fisher's use of "instantly" was similar to his use of "clearly" and "simply"; his readers learned to beware.

"The problem...derived from a sample of *n* pairs...may be solved... with the aid of certain very general conceptions derived from the geometry of *n* dimensional space." (Fisher, 1915, p. 508) Fisher remarked that the quantities under consideration "have, in fact, an exceedingly beautiful interpretation in generalized space....Considering first the space of *n* dimensions in which the variations of *x* are represented...." (p. 509) Fisher used an exclusively geometric argument to deduce that the point in question must lie within a certain sphere of $n-1$ dimensions. (In other words, a set of 10 observations was to be considered as the 10 coordinates of a point in Euclidean 10-space, on a sphere of 9 dimensions.)

Our interest in this work is Fisher's early and confident use of geometry to justify his mathematical statements. The $n-1$ dimensional spheres were in a paper published in 1915. Fisher used this same discussion 5 years later, in a paper on the mean error of a set of observations. (Fisher, 1920) In a note in his 1950 *Contributions to Mathematical Statistics*, Fisher commented on the earlier work. "Here the method of defining the sample by the coordinates of a point in Euclidean hyperspace was introduced...." (*CMS* 1.2a) Fisher recognized a bit of what he had done: Some years ago, the writer applied a novel method of geometrical representation to the problems of random sampling....(Fisher, 1921, p. 3) "In this method a sample is represented by a point in generalized space, the separate measurements being the coordinates of the point." (p. 5)

What was Fisher doing? Here he was not thinking of randomness as a requirement, but, rather, as a property. Randomness was something inherent in the process of sampling. Let us follow a bit down Fisher's path, to see how randomness can be connected to geometry. Note that, for Fisher, it was essential that a random sample be typical of any other random sample of the same size from the same population. So the geometrical point needed to be typical of any other possible point obtained in the same way.

The figure below is similar to Fisher's; it is the only one he put forth. According to Fisher, the point P would lie on a sphere of $n-1$ dimensions centered at M, the mean of the *n* observations. An element of surface of that sphere would then be discussed. In Fisher's view, all positions of P on that sphere were equally likely; that is, the occurrence of P in any element of surface was random (Figure 1).

In our case, Fisher's element of volume becomes a portion of the circle, a small arc. For Fisher, the meaning of all of this was that a point P was likely to be anywhere on the surface of the sphere, with equal probability; randomness was inherent.

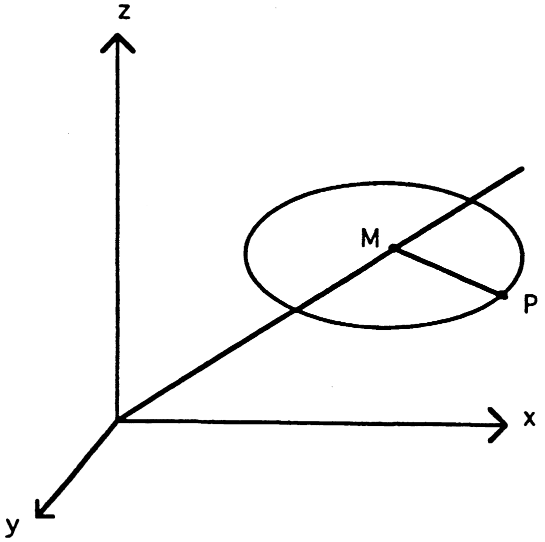

Figure 1. Sphere centered at M (Drawing suggested by Fisher, 1915, p. 509).

In a paper presented at the Royal Society in 1923, "The Influence of Rainfall on the Yield of Wheat at Rothamsted" (Fisher, 1925b), Fisher once again cited his geometric representation, this time with less emphasis on the aspects of Euclidean n-space and with a more concise statement of interpretation. "The mean of the observations thus specifies the origin, and the standard deviation specifies the length of the radius vector from the origin to the sample point." (Fisher, 1925b, p. 100)

The geometry above was arrived at by Fisher between 1915 and 1923 in order to validate and extend the work of Student, on small samples. I suggest that it was these considerations of small samples that led him to the concept of randomization in experimental design, probably in the later part of 1923. The statistical consideration of small samples is a thread running through a number of Fisher's published papers between 1915 and 1924 and in his voluminous correspondence.

Fisher's correspondence contains many discussions of small samples; in a letter to a plant pathologist:

> The mathematical problem solved by 'Student' in 1908 constituted a most important refinement of the theory of errors, the possibility of which had been previously overlooked by mathematical writers on this subject. He supplied for the first time a rigorously exact test of significance for the mean of a finite sample drawn from a normally

> distributed population. Perceiving the importance of what he had done, and being myself interested in the mathematical problems presented by the exact treatment of finite samples, I later extended his procedure....(Fisher, 1936 in Bennett, 1990, pp. 306–307)

In October of 1938, Fisher corresponded with Churchill Eisenhart of the Statistical Research Group at Columbia University, who was considering assembling an anthology of statistical readings.[4] Fisher was enthusiastic and suggested some of his own work. The first few papers in this category run from 1915 to 1923 and almost all of them touch on the mathematics of small samples.

One of Fisher' recommendations was his 1921 "On the 'Probable Error' of a Coefficient of Correlation Deduced from a Small Sample." The first section was entitled "The Curve of Random Sampling for 'Intraclass' Correlations"; in it Fisher once again used a geometric argument. "In this method a sample is represented by a point in generalized space, the separate measurements being the coordinates of the point." (Fisher, 1921, p. 5) The work on small samples and randomness continued. In "Note on Dr. Burnside's Recent Paper on Errors of Observation" (1923), Fisher used the geometric argument he had developed previously. "If we regard the observations $x_1, x_2, \dots x_n$ as coordinates in n-dimensional space, any set of observations will be represented by a single point...." (p. 656)

Fisher included his 1923 paper with Winifred Mackenzie, "Studies in Crop Variation. II. The Manurial Response of Different Potato Varieties," which used data from an experiment by Thomas Eden several years earlier. This paper is notable because the experimental design was systematic, but Fisher inserted a comment suggestive of randomizing. (This is the first such hint I have found, having examined Fisher's publications and his correspondence.) There were several paragraphs of description of the efforts taken to eliminate unwanted sources of variation. But in the analytical section of the paper, Fisher mentioned, "...if all the plots are undifferentiated, as if the numbers had been mixed up and written down in random order..." (p. 315)

Fisher wrote *Statistical Methods* in 1923 and the early part of 1924. In the preface, he was explicit about the need for mathematical statistics of small samples.

[4] This volume appeared in 1947, edited by Eisenhart and others. The full title was *Selected Techniques of Statistical Analysis for Scientific and Industrial Research and Production and Management Engineering*. It is a collection of seventeen articles, none of which is by Fisher.

> The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling *small sample problems on their own merits* does it seem possible to apply accurate tests to practical data. Such at least has been the aim of this book. (Fisher, 1925a, p. vii, emphasis added)

Thus, according to Fisher, the purpose of this book, his first, was to enable workers in laboratories and fields to evaluate their small sample data accurately. It was in this book that he first made explicit the requirement of randomization. In Chapter 8:

> [T]he experiment should yield not only a comparison of different manures, treatments, varieties, etc., but also a means of testing the significance of such differences as are observed....[I]f our plots have been chosen in any way according to a prearranged system...the systematic arrangement of our plots may have...features in common with the systematic variation of fertility, and thus the test of significance is wholly vitiated....[A] purely random arrangement of plots ensures that the experimental error calculated shall be an unbiased estimate of the errors actually present. (Fisher, 1925a, pp. 224–226)

According to Box, statisticians of the 1920's were very concerned with conditions of normality. Fisher seems to have been more concerned with independence of observations, or, rather, lack of independence. Clearly, observations from adjacent or nearby agricultural plots are not independent of each other. What was needed was an analytical technique, an analysis of variance, that did not require either observations that were known to be independent or observations drawn from a sample that was known to be normally distributed. Box suggested that Fisher must have had the insight to realize that if observations were collected with appropriate randomization, then "any observation was interchangeable with any other in the analytic expressions. On this basis, when the null hypothesis was true...the variance calculated for the group means was on the average identical with that within the groups....Thus randomized experiments could be analyzed *as if* the observations were roughly normally distributed and independent." (Box, 1978, pp. 148–149)

In the preface to *Statistical Methods* Fisher listed W. S. Gosset first among those he thanked. The Student-Fisher correspondence, one-sided though it is, provides us with a fascinating glimpse inside the minds of two of the most influential statisticians of the twentieth century. Of interest here will be the first few years. Both were concerned with the mathematics of small samples; Student[5] as the master brewer for Guinness, in Ireland, had a practical interest, and Fisher, it seems, earning his living as a consultant on agricultural experiments, enjoyed a mathematical challenge. Fisher's 1922 paper on the goodness of fit of regression formulae could well have been motivated by a question from Student. "I forgot to put up another problem to you in my last letter, that of the prob: error of partial correlation/regression coefficients for small samples." (Gosset, 1962, Ref. No. 6)

In the summer of 1924 Student agreed to read the proofs of Fisher's book, and by October he was returning them. Part of Fisher's summary of this period was as follows:

Ref. No.	Date of Letter	Contents
47.	14 July, 1924	Willing to read the proofs of my book which must have been available at this time;...
48.	17 July, 1924	My reply to earlier letter.
50.	20 Oct. 1924	Criticisms and advice after reading most of the proofs of "Statistical Methods," 1st edition; disagrees about randomization....
53.	4 Nov. 1924	Suggesting the omission of Example I as too difficult at the beginning of a book (I believe I did this in the 2nd edition); continues to disagree about randomization.

(Gosset 1962, vol. 5)

Student expressed part of his disagreement about randomization with good humor. Letter 50 contained 104 specific suggestions for changes in notation and presentation; Student prefaced them with, "I do not expect to convince you but I do not agree with your controlled randomness. You would want a large lunatic asylum for the operators who are apt to make mistakes enough even at present." (Gosset, 1962, Ref. No. 50) Student did not adopt Fisher's view of the need for randomization; Fisher's summary of Letter 68 in November of 1925 stated, "On some experiments trying randomization against systematic designs. He is still, as he remained later, adverse to randomization in field experimentation."

[5] Student, or W. S. Gosset, as he signed his letters, was 14 years older than Fisher. Gosset was married with children, had many interests outside the brewery, and had an engaging sense of humor. In a letter to Fisher he described a calculating machine at the brewery as having been used by Noah for quantitative calculations before his voyage and then bartered to the brewery in exchange for a barrel of porter.

Statistical Methods, examined today, appears to be just another statistics book, but it was the first to advocate randomization rather than systematic experimental design, and it was the first to explicate analysis of variance, which Fisher had originated only a few years before. Because of that, and the fact that Fisher wrote for the reader who was a practicing experimenter, the book went through many editions and several languages.

The last section of *Statistical Method*, Section 48, is titled "Technique of Plot Experimentation."

> The first requirement which governs all well-planned experiments is that the experiment should yield not only a comparison of different manures, treatments, varieties, etc., but also a means of testing the significance of such differences as are observed. (Fisher, 1925a, p. 224)

> For our test of significance to be valid the difference in fertility between plots chosen as parallels must be truly representative of the differences between plots with different treatments; and we cannot assume that this is the case if our plots have been chosen in any way according to a pre-arranged system; for the systematic arrangement of our plots may have...features in common with the systematic variation of fertility, and thus the test of significance is wholly vitiated. (Fisher, 1925a, p. 224)

In other words, if the systematic scheme of the researcher happens to correlate in some way with the (perhaps unobservable) systematic non-uniformity of the field, then any test of significance will be meaningless.

What now? Fisher had the answer. "The direct way of overcoming this difficulty is to arrange the plots wholly at random." (p. 225) And, lest the reader interpret "random" as meaning "according to whim," Fisher immediately launched into an example where twenty plots of land were to be used to test five treatments; the arrangement was to be arrived at "by shuffling 20 cards thoroughly and setting them out in order." (p. 225) Here, briefly, Fisher led his reader through randomized blocks and Latin squares.[6] In advocating randomization over

[6] In a Latin square, each treatment occurs once in each row and once in each column. Here are some 4 × 4 Latin squares

ABDC	CBAD	DCAB
BCAD	BCDA	ABCD
CDBA	ADBC	CDBA
DACB	DACB	BADC

systematic arrangements, he pointed out that any systematic arrangement had to be made in accordance with assumptions that the researcher had as to soil heterogeneity, fertility, drainage, etc. and that these assumptions might or might not be valid; thus the standard error could be calculated in any one of several ways, giving various results according to the assumptions of the researcher. Randomization eliminated the need for these assumptions.

The nine concluding pages of *Statistical Methods* became the basis for *The Design of Experiments* eleven years later. According to Box (1978), *Statistical Methods* "did not receive a single good review." (p. 130) But the book, after starting slowly, became more and more widely known and used. A second edition was published in 1928 and a third in 1930. *Statistical Methods* was eagerly received by those doing research in the biological sciences, where sample sizes were customarily small and older statistical methods were inappropriate.

Between 1926 and 1930 Fisher published four papers in which randomization featured prominently.[7] The need for randomization had become at least widely known within the agricultural community, if not completely accepted, by the early 1930's. According to Box (1978) it was Fisher's teaching of statistical methods at two colleges in 1930 followed by his 1931 summer at Ames, Iowa, and the notes for these lectures, that became the basis for Fisher's next book, *The Design of Experiments.* At Ames, at the Iowa State College of Agriculture, now Iowa State University, Fisher gave lectures and seminars on statistical methods, including the design of experiments, and also on genetics.

The Design of Experiments (1935) is exposition and instruction; it is also a carefully constructed sales effort, an effort by Fisher to sell his theories and accompanying methods to both highly educated researchers and minimally educated agricultural workers. He appealed to the use of everyday simple items; interspersed with these suggestions were mathematical discussions of the utmost rigor. After an introduction, *Design* began with the charming tale of a lady who declared she could tell by tasting whether tea or milk was placed in her cup first. Using the familiar activity of tea tasting, Fisher discussed various possible experimental designs, involving null hypothesis, randomization and repetition.

[7] 1926: "The Arrangement of Field Experiments"; 1927: (with Eden) "Studies in Crop Variation. IV. The Experimental Determination of the Value of Top Dressings with Cereals"; 1929: (with Eden) "Studies in Crop Variation. VI. Experiments on the Response of the Potato to Potash and Nitrogen"; 1930: (with Wishart) "The Arrangement of Field Experiments and the Statistical Reduction of the Results"; 1931: "The Techniques of Field Experiments."

This episode has gained prominence in the folklore of the history of statistics; when Fisher first described it in *Design*, most readers assumed it to be explanatory fiction. But, according to David Salzburg (2001), in *The Lady Tasting Tea*, the event occurred one summer afternoon at Cambridge University in the late 1920's. Salzburg was a student of H. Fairfield Smith, who claimed to have been a participant that day. Fisher and the other assorted academicians had the fun of designing an experiment, on the spot, to see if the lady could indeed tell whether tea or milk went into the cup first. Eight cups of tea were prepared, four of each type, carefully randomized (the method was unspecified), and then offered to her. According to Smith, she got every one correct. (Salzburg, 2001) (In a differing version, Box (1978, p. 134) has written that the event occurred at Rothamsted, early in Fisher's years there. The lady was Muriel Bristol, and another participant was William Roach, a colleague of Fisher's. According to Roach, who later married Miss Bristol, she got enough of the cups right to prove her point.)

Having started with the basic experimental design of four cups of tea of each type, Fisher then instructed upon tests of significance, the null hypothesis, and the role of randomization. He included a discussion of the sensitivity of the experiment and the effects of repeating the experiment, of using more cups of tea, or of not requiring that an equal number of cups be prepared each way.

The next chapter, "A Historical Experiment on Growth Rate," was devoted entirely to Charles Darwin's experimental design in Darwin's *The Effects of Cross and Self-Fertilization in the Vegetable Kingdom.* Darwin wanted to verify the often repeated claim that cross fertilized plants would be superior to those self fertilized. He had established fifteen pairs of plants to be compared as to height; each of fifteen crossed plants was assigned to a self-bred specimen. Care was taken that all were of the same age; they were then planted in pots, each pair in the same pot, to eliminate differences of soil, moisture, light, etc. At maturation, each plant was measured and the height of the self-fertilized plant was subtracted from that of the cross-bred. This resulted in thirteen positive differences and two negative; in thirteen cases the cross-bred plant was taller than its self-fertilized companion.

Fisher led his reader through an analysis of variance and a test of significance using Student's *t*-test; this presentation is astonishing for being lucid and brief – less than two pages. After disagreeing with parts of Galton's statistical methods, as related by Darwin, Fisher then discussed error and randomization.

> Having decided that...our estimate of the error of the average difference must be based upon the discrepancies between the differences actually observed, we must next enquire what precautions are needed...to guarantee that such an estimate shall be a valid one; that is to say that the very same causes that produce our real error shall also contribute the materials for computing an estimate of it....(Fisher, 1935, p. 46)

> In the experiment under consideration, the sole source of the experimental error in the average of our fifteen differences lies in the differences in soil fertility, illumination, evaporation, etc., which makes the site of each crossed plant more or less favourable to growth than the site assigned to the corresponding self-fertilised plant...[W]hen the fifteen pairs of sites have been chosen...we then assign at random, as by tossing a coin, which site shall be occupied by the crossed and which by the self-fertilised plant... (1935, pp. 47–48)

> Randomisation, properly carried out, in which each pair of plants is assigned their positions independently at random, ensures that the estimates of error will take proper care of all such causes of different growth rates, and relieves the experimenter from the anxiety of considering and estimating the magnitude of the innumerable causes by which his data may be disturbed. The one flaw in Darwin's procedure was the absence of randomisation. (1935, p. 9)

In other words, according to Fisher, Darwin's experimental design was a good one; the only thing lacking was randomization, perhaps by flipping a coin, to determine which of each pair went where (within each pair, whether to place the self-fertilized plant on the left or the right). Subsequent chapters of *Design* discussed randomized blocks, Latin squares, and several other designs of complex experiments.

The Design of Experiments was welcomed by some in the statistical community and by the research community. It seems that this difference in reception can be attributed to two things, both Fisher's stature and reputation by this time, and to the fact that some of Fisher's ideas that had seemed so novel 10 years before had by now achieved a level of wide familiarity, if not complete acceptance.

For Fisher, much had happened in the 10 year interval between the publications of *Statistical Methods* and *Design*. Now, in 1935, he was R. A. Fisher, Sc.D., F.R.S. That is, he was now Dr. Fisher, holding a degree from Cambridge University, and a fellow of the Royal Society of

London. He was now Galton Professor of Eugenics at University College, London, having left Rothamsted in 1933. And, while it is not relevant to our topic of randomization, we note that in 1930 he had published *The Genetical Theory of Natural Selection*, a major contribution to the synthesis of Darwinian natural selection with Mendelian heredity.

A number of statisticians and agriculturalists were reluctant to give up systematic experimental designs for the seemingly less organized ones resulting from randomization. But as Fisher's methods became known, laboratories and organizations all over the world began writing to him for statistical and design advice, as his correspondence attests. They also began sending their personnel for training, to work with Fisher. At Rothamsted these were known as "voluntary workers." According to Box (1978) their period of residence varied from as little as three weeks to as long as 3 years, supported by their home institution. "In 1928 there were three, in 1929 nine, and in 1930 thirteen..." (Box, 1978, p. 157) This pattern continued for the next several years, even after Fisher left Rothamsted in 1933; voluntary workers were there with him at the Galton Laboratory. The Appendix is a reconstruction of the personnel roster of these "voluntary workers" with Fisher during his Rothamsted years. They came from far and wide and returned there, taking with them Fisherian statistical methods and Fisherian experimental design. Even a cursory look at their later publications makes it clear that they actually put into practice what they learned at Rothamsted.

The notes of Gavin Ross, now retired from the Statistical Department at Rothamsted, on Fisher and Fisher's associates have provided a valuable framework. Other sources were the *Records of the Rothamsted Staff, Harpenden* and the series of official reports published by Rothamsted Experimental Station. In the *Report 1927–1928* was this description.

> The activities of Rothamsted, however, are not confined to the British Islands, but are gradually spreading out to the Empire and other countries abroad.... More and more workers are coming from the overseas Dominions to carry on their studies at Rothamsted. None but University graduates are eligible, and most are, or are about to be, on the staffs of Government or other Agricultural Departments: men who will become leaders in the agricultural communities of their respective countries. (p. 19)

Fisher's Statistical Laboratory started with only himself and Winifred Mackenzie as his assistant in 1919. Over the years it grew, so that by

1928 there were two Assistant Statisticians (J. Wishart and J. O. Irwin) and three Assistant Computers (A. D. Dunkley, Florence Pennells, Alice Kingman)[8] in addition to others, those to whom the above quotation applies. Some were referred to, officially, as Post-Graduate Research workers, some as Temporary Workers, some as Short Period Workers and some seemed to have no official title; it seems that Box has included all of these in her designation "voluntary workers." All of these were either on research scholarships or were being supported by their home institutions. In the Appendix are who they were, when they were at Rothamsted, and, most important, where they went after their study with Fisher. The 35 identified so far traveled to Australia (including Tasmania), Brazil, Canada, Ceylon, Germany, Greece, Holland, India, Ireland, Sweden, Tanganyika, Trinidad, Uganda, and the United States. The Appendix includes only those who were officially attached to the Statistical Laboratory.

In addition to this group of voluntary workers there were several of the Rothamsted staff in other departments who published papers advocating Fisher's methods and who moved on from Rothamsted to locations around the world. T. Eden was in the Field Experiments Department from 1921 to 1927; he published several papers with Fisher. Upon leaving Rothamsted he became a chemist at the Tea Research Institute of Ceylon; he came back to Rothamsted in 1932 to spend his leave with Fisher. J. K. Basu, in the Physics and Chemistry Laboratories, 1927–1930, went to the Imperial Institute of Agricultural Research, Pusa, India. A. R. Clapham, Plant Physiologist, 1928–1930, became Demonstrator at the School of Botany, Oxford University. E. J. Maskell, in the Field Experiments Department from 1924 to 1926, became Assistant Plant Physiologist at the College of Tropical Agriculture, Trinidad. And, two of Fisher's own, the Assistant Statisticians, Irwin and Wishart, moved on. J. O. Irwin, 1928–1930, became assistant to the chair of the Statistics Committee of the Medical Research Council. J. Wishart, 1927–1931, became Reader in Statistics, Cambridge University. Again, examination of their later publications makes Fisher's influence clear.

The stream of visitors continued, at the Galton Laboratory after Fisher moved there in 1933. In his 1950 address, "The Fisherian Revolution in Methods of Experimentation," W. J. Youden said, "Last summer I examined the record of the Galton Laboratory of the workers

[8] In all of the Rothamsted records cited here, the *Reports* and the *Records of the Rothamsted Staff*, it seems to have been the custom to use the first names of women but only the initials of men.

who studied with Fisher in the period 1933–1944. There are over fifty names, from more than a score of countries, working in a score of different scientific fields." (Youden, 1951, p. 49)

In the late 1920's and the 1930's, Fisher's disciples were using and instructing on his methods, including randomized experimental design, worldwide. Writing in 1939 on the occasion of a tribute to A. Daniel Hall, Russell summarized the impact of Fisher's experimental designs.

> The new methods have proved so much more useful than the old that they are now used almost all over the Empire and in parts of the United States. From Rothamsted we are advising on, or in some cases actually conducting, experiments on rubber in Malaya, tea in India, cotton in the Sudan, and oil palm in West Africa, while our old research workers are using the methods for experiments on sugar-cane and rice in India, tea in Ceylon, to say nothing of numerous experiments in the United Kingdom. The methods are spreading, and are constantly being improved. (Russell, 1939, p. 170)

Another indication of Fisher's enunciation of randomization in field experiments can be seen in a definite shift within papers published in English journals, such as the *Journal of Agricultural Science*. Fisher's methods, including randomization, appeared first in the 1920's in the papers of those who were or had been at Rothamsted, complete with explanations of randomized blocks and Latin squares. By the 1930's a researcher could report that the arrangement was a Latin square and proceed to describe the rest of the experiment; no explanation of Latin square was needed.

The Times of London carried Fisher's obituary on July 31, 1962. In describing Fisher's years at Rothamsted:

> Here Fisher began the remarkable series of statistical investigations which led to the techniques described in *Statistical Methods for Research Workers* (1925, 10 th edition 1946), *The Design of Experiments* (1935, fifth edition 1949) and *Statistical Tables* (published with Frank Yates, 1938, third edition 1947). These works revolutionized agricultural research; for they described the methods, now used the world over, for evaluating the results of small sample experiments and for so laying our experimental trials as to minimize the disturbances due to heterogeneity of soils and the unavoidable irregularities of biological material. ("Obituary – Sir Ronald Fisher")

Conclusion

In his address to the American Statistical Association in 1950, W. J. Youden summarized the impact of Fisher's methods.

> In the past 25 years, a far-reaching and virtually revolutionary development in the technique of experimentation has been taking place. In many laboratories men began to ask what are the requirements that must be met in order to permit valid conclusions from experiments. The analysis of variance...showed experimenters the importance of placing the comparisons allotted to treatments on the same basis as the comparisons used for the estimate of error. The process of random assignment or arrangement was seen to be an essential and easy method of giving the same opportunity to treatment and error methods....The analysis of variance for the first time provided a sound statistical technique to accompany experimental arrangements such as the replicated block and the Latin square.... (Youden, 1951, pp. 48–54)

Fisher's requirement of randomization initiated a methodological revolution in experimental research. In his view randomization served two purposes: it eliminated bias and it enabled a valid test of significance. The bias of concern could have originated in the experimental material, as in differences in soil heterogeneity or, in social experiments, differences in school populations; alternately, the bias could have been with the researcher, who, perhaps inadvertently might have placed some treatment in an advantageous position relative to other treatments. Using a strict randomizing device, whether cards or a random number table, or something similar, along with sufficient replication, would insure that no treatment had any advantage or disadvantage.

Fisher's second reason for requiring randomization was that it made possible a valid test of significance, a measure of how likely the statistic calculated from the data was a good estimate of the property being considered. Because treatments were assigned to units using a randomizing device, the rules of probability could be invoked, and the likelihood that the result was *not* due to chance could be calculated.

In his early work on the statistics of small samples, stimulated by the work of Student, Fisher's own view of the randomness inherent in sampling was a geometric one; it began with considering n observations as the n coordinates of a point in Euclidean n-space. It was his consideration of the statistical methods used in small samples, and the important role that randomness plays there, that, this paper suggests, led Fisher to require randomization in the layout of an experiment.

In sum, Fisher first proposed randomization in his *Statistical Methods for Research Workers*, published in 1925 while he was working at Rothamsted Agricultural Station in Harpenden, England. From there the word spread about this way of doing research, along with Fisherian statistical methods and principles of experimental design. First came English agriculturalists, and later research workers and statisticians from many disciplines and many nationalities, to visit Rothamsted and to study under Fisher; they returned home and wrote and taught Fisher's methods. By the 1950's Fisher's requirement of randomization, his principles of experimental design, and his accompanying methods of statistical analysis could be found in statistics books in many countries around the world. His randomized experimental designs changed the way that many experiments are carried out and his principles of experimental design are now basic to scientific experiment.

Acknowledgments

I thank Lindley Darden for her patience and her continued support, and the editorial staff and anonymous reviewers at the *Journal for the History of Biol*ogy for their excellent comments. This paper is based upon my dissertation, *R. A. Fisher and Randomized Experimental Design*, which was supported by National Science Foundation Doctoral Dissertation Research Grant No. 0091782.

Appendix. Fisher's voluntary workers at Rothamsted

Person	When	Home	Later institution
E. Somerfield	1923	Dublin	Messrs. A. Guinness & Sons, Ltd.
L. H. C. Tippett	1924–1926	British Cotton Ind. Research Assn.	Shirley Institute, Manchester
Ansell, P. R.	1925		
T. N. Hoblyn	1926–1927		East Malling Fruit Research Institute
Balmukand	1927–1928	Punjab U.	Punjab Agricultural College, India
W. Boehme	1928	Germany	
J. B. Hutchinson	1928–1929	Trinidad	Empire Cotton Res. Sta., Trinidad; 1933, Inst. of Plant Industry, Indore, Central India

Appendix. Continued

Person	When	Home	Later institution
T. W. Simpson	1928–1999		
Howard Hotelling	1929	California	Stanford University
H. G. Sanders	1929		
Frances E. Allan	1929–1930	Melbourne U.	Statistician, Council for Sci. and Industrial Research, Canberra
Edgar Anderson	1929–1930	Washington University	Missouri Botanic Garden
R. J. Kalamkar	1929–1932	Nagpur U., India	Asst. Meteorologist, Poona, India
Basil Christidis	1930		Plant Breeding Station, Greek Ministry of Ag., Salonika
Alcides Franco	1930		Ministry of Ag., Rio de Janeiro
C. H. Goulden	1930		Dominion Rust Research Lab., Manitoba Ag. College, Winnipeg
C. H. N. Jackson	1930	Tanganyika	Entomologist, Dept. of Tsetse Research, Tanganyika Territory
J. W. Hopkins	1930–1931	U. of Alberta	Nat. Research Council, Ottawa, Canada
F. R. Immer	1930–1931	St. Paul, Minn.	Cereal Geneticist, USDA, St. Paul, Minn.
A. W. R. Joachim	1930–1931		Dept. of Agriculture, Ceylon
A. L. Murray	1930–1931	Dublin	Messrs. A. Guinness & Sons, Ltd.
S. H. Justensen	1931		The University, Wageningen, Holland
R. Summerby	1931		Agronomy Dept., Macdonald College, Quebec
H. J. Buchanan-Wollaston	1931–1932		Fisheries Laboratories, Lowestoft
J. Rasmussen	1931–1932	Sweden	Astrund, Sweden
R. O. Iliffe	1932	Ag. Res. Institute, Coimbatore, India	
S. A. Stouffer	1932	U. of Chicago	Department of Sociology, U. of Wisconsin, Madison
C. Stuart Christian	1932–1933		Queensland U., Brisbane
R. S. Koshal	1932–1933		Tech. Research, Indian Central Cotton Committee, Bombay
A. Bigot	1933	Agricultural H. S., Wageningen, Holland	
A. E. Brandt	1933	Iowa State Col., Ames, IA	Iowa State College of Agriculture Ames, IA
H. L. G. Milne	1933	Dept. of Ag., Entebbe, Uganda	East African Ag. Research Station, Tanganyika Territory
R. A. Scott	1933	Dept. of Agriculture, Launceston, Tasmania	
I. Zacopanay	1933		
A. V. Coombs	1933–1934		Imperial Chem. Industries, Colombo, Ceylon

References

Bennett, J. H. (eds.) 1990. *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher*. Oxford: Clarendon Press.

Box, George E. P. 1976. "Science and Statistics." *Journal of the American Statistical Association* 356: 791–799.

Box, Joan Fisher. 1978. *R. A. Fisher: The Life of a Scientist*. New York: John Wiley and Sons.

—— 1980. "R. A. Fisher and the Design of Experiments, 1922–1926." *The American Statistician* 34: 1–7.

Darwin, Charles. 1876. *The Effects of Cross- and Self-Fertilisation in the Vegetable Kingdom*. London: John Murray.

Dehue, Trudy. 1997. "Deception, Efficiency, and Random Groups: Psychology and the Gradual Origination of the Random Group Design." *Isis* 88: 653–673.

Eden, Thomas and Ronald A, Fisher. 1927. "Studies in Crop Variation. IV. The Experimental Determination of the Value of Top Dressing with Cereals." *Journal of Agricultural Science* 17: 548–562.

—— 1929. "Studies in Crop Variation. VI. Experiments on the Response of the Potato to Potash and Nitrogen." *Journal of Agricultural Science* 19: 201–213.

Eisenhart, Churchill, Millard W. Hastay and W. Allen Wallis (eds.). 1947. *Selected Techniques of Statistical Analysis for Scientific and Industrial Research and Production and Management Engineering*. New York: McGraw Hill.

Fechner, Gustav Theodor. 1860 [1966]. *Elemente der Psychophysik*. Trans. by Helmut E. Adler. *Elements of Psychophysics*. New York: Holt, Rinehart and Winston, Inc.

Fisher, Ronald A. 1915. "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population." *Biometrika* 10: 507–521.

—— 1918. "The Causes of Human Variability." *Eugenics Review* 10: 213–220.

—— 1920. "A Mathematical Examination of the Methods of Determining the Accuracy of an Observation by the Mean Error, and by the Mean Square Error." *Monthly Notices of the Royal Astronomical Society* 80: 758–770.

—— 1921. "On the 'Probable Error' of a Coefficient of Correlation Deduced from a Small Sample." *Metron* 1: 3–32.

—— 1922. "The Goodness of Fit Regression Formulae and the Distribution of Regression Coefficients." *Journal of the Royal Statistical Society* 85: 597–612.

—— 1923. "Note on Dr. Burnside's Recent Paper on Errors of Observation," *Proceedings of the Cambridge Philosophical Society* 21: 655–658.

—— 1925a. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

—— 1925b. "The Influence of Rainfall on the Yield of Wheat at Rothamsted." *Philosophical Transactions of the Royal Society of London. Series B*. 213: 89–142.

—— 1926. "The Arrangement of Field Experiments." *Journal of the Ministry of Agriculture* 33: 503–515.

—— 1930. *The Genetical Theory of Natural Selection*. Oxford: University Press.

—— 1931. "Principles of Field Experimentation in Relation to the Statistical Interpretation of the Results," In: *The Technique of Field Experiments*. Harpenden: Rothamsted Experimental Station, pp. 11–13.

—— 1935. *The Design of Experiments*. Edinburgh: Oliver and Boyd.

—— 1950. *Contributions to Mathematical Statistics*. New York: Wiley.

—— 1971–1974. *Collected Papers of R. A. Fisher*. J. H. Bennett (ed.), Adelaide, Australia: University of Adelaide, Vol. 1–5.

Fisher, Ronald A. and Winifred A, Mackenzie. 1923. "Studies in Crop Variation. II. The Manurial Response of Different Potato Varieties." *Journal of Agricultural Science* 13: 311–320.

Fisher, Ronald A. and Wishart, J. 1930. *The Arrangement of Field Experiments and the Statistical Reduction of the Results*. London: Imperial Bureau of Soil Science.

Gosset, William S. 1962. *Letters from W. S. Gosset to R. A. Fisher, 1915–1936*, Vol. 5. Dublin: L. McMullen.

Hacking, Ian. 1988. "Telepathy: Origins of Randomization in Experimental Design." *Isis* 79: 427–451.

Hall, Nancy S. 2002a. "The R. A. Fisher Collection at Adelaide University." *Mendel Newsletter*, pp. 2–6.

—— 2002b. *R. A. Fisher and Randomized Experimental Design*. Ph.D. dissertation. College Park, MD: University of Maryland.

Levi, Isaac. 1983. "Direct Inference and Randomization." D. Asquith Peter and Thomas Nickles (eds.), *PSA 1982*, Vol 2. East Lansing, MI: Philosophy of Science Association, pp. 447–463.

McCall William, A. 1923. *How to Experiment in Education*. New York: Macmillan.

Mercer, W. B. and Hall, A.D. 1911. "The Experimental Error of Field Trials." *Journal of Agricultural Science* 4: 107–132.

"Obituary - Sir Ronald Fisher." 31 July 1962. London: *The Times*.

Owen, D. B. (ed.). 1976. *On the History of Statistics and Probability*. New York: Marcel Dekker, Inc.

Peirce, Charles S. and Joseph Jastrow. 1885. "On Small Differences in Sensation." *Memoirs of the National Academy of Sciences* 3: 73–83. Reprinted in Arthur W. Burks (ed.). 1958. *Collected Papers of Charles Sanders Peirce*. Cambridge: Harvard University Press. Vol. 7, pp. 13–34.

Preece, D. A. 1990. "R. A. Fisher and Experimental Design: A Review." *Biometrics* 46: 925–935.

Ross, Gavin. 2002. Notes: "Fisher and Rothamsted." Given by him to the author.

Rothamsted Experimental Station. 1929–1935. *Records of the Rothamsted Staff, Harpenden*. No. 1–5. St. Albans: Gainsborough Press.

—— 1921–35. Report 1918–20, 1921–22,23–24,25–26,27–28,1929,1930,1931,1932, 1933, 1934. Harpenden: D. J. Jeffery.

—— 1914. *Rothamsted Memoirs on Agricultural Science*. Harpenden: D. J. Jeffrey.

Russell, E. John. 1939. "Soil Science in England 1894–1938," in *Agriculture in the Twentieth Century: Essays on Research, Practice and Organization to be Presented to Sir Daniel Hall*. Oxford: Clarendon Press, pp. 163–191.

—— 1956. *The Land Called Me: An Autobiography*. London: George Allen and Unwin Ltd.

—— 1966. *A History of Agricultural Science in Great Britain*. London: George Allen and Unwin Ltd.

Salzburg, David. 2001. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. New York: W. H. Freeman and Company.

Stigler, Stephen M. 1978. "Mathematical Statistics in the Early States." *Annals of Statistics* 6: 239–265.

—— 1986. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, MA: Harvard University Press.

Student. 1908a. "The Probable Error of a Mean." *Biometrika* 6: 1–25.

—— 1908b. "Probable Error of a Correlation Coefficient." *Biometrika* 6: 302–310.

Tankard, James W. 1984. *The Statistical Pioneers.* Cambridge, MA: Schenkman Publishing Company.

The October Meeting of the National Academy of Sciences. 1884. *Science* 90(4): 396–397.

Wood, T. B. and Stratton, F. J. M. 1910. "The Interpretation of Experimental Results." *Journal of Agricultural Science* 3: 417–440.

Youden, W. J. 1951. "The Fisherian Revolution in Methods of Experimentation." *Journal of the American Statistical Association* 46: 47–50.

Journal of Economic Literature 48 (June 2010): 424–455
http:www.aeaweb.org/articles.php?doi=10.1257/jel.48.2.424

Instruments, Randomization, and Learning about Development

Angus Deaton*

There is currently much debate about the effectiveness of foreign aid and about what kind of projects can engender economic development. There is skepticism about the ability of econometric analysis to resolve these issues or of development agencies to learn from their own experience. In response, there is increasing use in development economics of randomized controlled trials (RCTs) to accumulate credible knowledge of what works, without overreliance on questionable theory or statistical methods. When RCTs are not possible, the proponents of these methods advocate quasi-randomization through instrumental variable (IV) techniques or natural experiments. I argue that many of these applications are unlikely to recover quantities that are useful for policy or understanding: two key issues are the misunderstanding of exogeneity and the handling of heterogeneity. I illustrate from the literature on aid and growth. Actual randomization faces similar problems as does quasi-randomization, notwithstanding rhetoric to the contrary. I argue that experiments have no special ability to produce more credible knowledge than other methods, and that actual experiments are frequently subject to practical problems that undermine any claims to statistical or epistemic superiority. I illustrate using prominent experiments in development and elsewhere. As with IV methods, RCT-based evaluation of projects, without guidance from an understanding of underlying mechanisms, is unlikely to lead to scientific progress in the understanding of economic development. I welcome recent trends in development experimentation away from the evaluation of projects and toward the evaluation of theoretical mechanisms. (JEL C21, F35, O19*)*

* Deaton: Princeton University. This is a lightly revised and updated version of "Instruments of Development: Randomization in the Tropics and the Search for the Elusive Keys to Economic Development," originally given as the Keynes Lecture, British Academy, October 9, 2008, and published in the Proceedings of the British Academy 102, 2008 Lectures, © The British Academy 2009. I am grateful to the British Academy for permission to reprint the unchanged material. I have clarified some points in response to Imbens (2009), but have not changed anything of substance and stand by my original arguments. For helpful discussions and comments on earlier versions, I am grateful to Abhijit Banerjee, Tim Besley, Richard Blundell, Ron Brookmayer, David Card, Anne Case, Hank Farber, Bill Easterly, Erwin Diewert, Jean Drèze, Esther Duflo, Ernst Fehr, Don Green, Jim Heckman, Ori Heffetz, Karla Hoff, Bo Honoré, Štěpán Jurajda, Dean Karlan, Aart Kraay, Michael Kremer, David Lee, Steve Levitt, Winston Lin, John List, David McKenzie, Margaret McMillan, Costas Meghir, Branko Milanovic, Chris Paxson, Franco Peracchi, Dani Rodrik, Sam Schulhofer-Wohl, Jonathan Skinner, Jesse Rothstein, Rob Sampson, Burt Singer, Finn Tarp, Alessandro Tarozzi, Chris Udry, Gerard van den Berg, Eric Verhoogen, Susan Watkins, Frank Wolak, and John Worrall. I acknowledge a particular intellectual debt to Nancy Cartwright, who has discussed these issues patiently with

1. *Introduction*

The effectiveness of development assistance is a topic of great public interest. Much of the public debate among noneconomists takes it for granted that, if the funds were made available, poverty would be eliminated (Thomas Pogge 2005; Peter Singer 2004) and at least some economists agree (Jeffrey D. Sachs 2005, 2008). Others, most notably William Easterly (2006, 2009), are deeply skeptical, a position that has been forcefully argued at least since P. T. Bauer (1971, 1981). Few academic economists or political scientists agree with Sachs's views, but there is a wide range of intermediate positions, well assembled by Easterly (2008). The debate runs the gamut from the macro—can foreign assistance raise growth rates and eliminate poverty?—to the micro—what sorts of projects are likely to be effective?—should aid focus on electricity and roads, or on the provision of schools and clinics or vaccination campaigns? Here I shall be concerned with both the macro and micro kinds of assistance. I shall have very little to say about what actually works and what does not—but it is clear from the literature that we do not know. Instead, my main concern is with how we should go about finding out whether and how assistance works and with methods for gathering evidence and learning from it in a scientific way that has some hope of leading to the progressive accumulation of useful knowledge about development. I am not an econometrician, but I believe that econometric methodology needs to be assessed, not only by methodologists, but by those who are concerned with the substance of the issue. Only they (we) are in a position to tell when something has gone wrong with the application of econometric methods, not because they are incorrect given their assumptions, but because their assumptions do not apply, or because they are incorrectly conceived for the problem at hand. Or at least that is my excuse for meddling in these matters.

Any analysis of the extent to which foreign aid has increased economic growth in recipient countries immediately confronts the familiar problem of simultaneous causality; the effect of aid on growth, if any, will be disguised by effects running in the opposite direction, from poor economic performance to compensatory or humanitarian aid. It is not obvious how to disentangle these effects, and some have argued that the question is unanswerable and that econometric studies of it should be abandoned. Certainly, the econometric studies that use international evidence to examine aid effectiveness currently have low professional status. Yet it cannot be right to give up on the issue. There is no general or public understanding that nothing can be said, and to give up the econometric analysis is simply to abandon precise statements for loose and unconstrained histories of episodes selected to support the position of the speaker.

The analysis of aid effectiveness typically uses cross-country growth regressions with the simultaneity between aid and growth dealt with using instrumental variable methods. I shall argue in the next section that there has been a good deal of misunderstanding in the literature about the use of instrumental variables. Econometric analysis has changed its focus over the years, away from the analysis of models derived from theory toward much looser specifications that are statistical representations of program evaluation. With

me for several years, whose own work on causality has greatly influenced me, and who pointed me toward other important work; to Jim Heckman, who has long thought deeply about these issues, and many of whose views are recounted here; and to the late David Freedman, who consistently and effectively fought against the (mis)use of technique as a substitute for substance and thought. None of which removes the need for the usual disclaimer, that the views expressed here are entirely my own. I acknowledge financial support from NIA through grant P01 AG05842–14 to the National Bureau of Economic Research.

this shift, instrumental variables have moved from being solutions to a well-defined problem of inference to being devices that induce quasi-randomization. Old and new understandings of instruments coexist, leading to errors, misunderstandings, and confusion, as well as unfortunate and unnecessary rhetorical barriers between disciplines working on the same problems. These abuses of technique have contributed to a general skepticism about the ability of econometric analysis to answer these big questions.

A similar state of affairs exists in the microeconomic area, in the analysis of the effectiveness of individual programs and projects, such as the construction of infrastructure—dams, roads, water supply, electricity—and in the delivery of services—education, health, or policing. There is frustration with aid organizations, particularly the World Bank, for allegedly failing to learn from its projects and to build up a systematic catalog of what works and what does not. In addition, some of the skepticism about macro econometrics extends to micro econometrics, so that there has been a movement away from such methods and toward randomized controlled trials. According to Esther Duflo, one of the leaders of the new movement in development, "Creating a culture in which rigorous randomized evaluations are promoted, encouraged, and financed has the potential to revolutionize social policy during the 21st century, just as randomized trials revolutionized medicine during the 20th," this from a 2004 *Lancet* editorial headed "The World Bank is finally embracing science."

In section 4 of this paper, I shall argue that, *under ideal circumstances*, randomized evaluations of projects are useful for obtaining a convincing estimate of the average effect of a program or project. The price for this success is a focus that is too narrow and too local to tell us "what works" in development, to design policy, or to advance scientific knowledge about development processes. Project evaluations, whether using randomized controlled trials or nonexperimental methods, are unlikely to disclose the secrets of development nor, unless they are guided by theory that is itself open to revision, are they likely to be the basis for a cumulative research program that might lead to a better understanding of development. This argument applies a fortiori to instrumental variables strategies that are aimed at generating quasi-experiments; the value of econometric methods cannot and should not be assessed by how closely they approximate randomized controlled trials. Following Nancy Cartwright (2007a, 2007b), I argue that evidence from randomized controlled trials can have no special priority. Randomization is not a gold standard because "there is no gold standard" Cartwright (2007a.) Randomized controlled trials cannot automatically trump other evidence, they do not occupy any special place in some hierarchy of evidence, nor does it make sense to refer to them as "hard" while other methods are "soft." These rhetorical devices are just that; metaphor is not argument, nor does endless repetition make it so.

More positively, I shall argue that the analysis of projects needs to be refocused toward the investigation of potentially generalizable mechanisms that explain why and in what contexts projects can be expected to work. The best of the experimental work in development economics already does so because its practitioners are too talented to be bound by their own methodological prescriptions. Yet there would be much to be said for doing so more openly. I concur with Ray Pawson and Nick Tilley (1997), who argue that thirty years of project evaluation in sociology, education, and criminology was largely unsuccessful because it focused on *whether* projects worked instead of on *why* they worked. In economics, warnings along the same lines have been repeatedly given by James J. Heckman (see particularly

Heckman 1992 and Heckman and Jeffrey A. Smith 1995), and much of what I have to say is a recapitulation of his arguments.

The paper is organized as follows. Section 2 lays out some econometric preliminaries concerning instrumental variables and the vexed question of exogeneity. Section 3 is about aid and growth. Section 4 is about randomized controlled trials. Section 5 is about using empirical evidence and where we should go now.

2. *Instruments, Identification, and the Meaning of Exogeneity*

It is useful to begin with a simple and familiar econometric model that I can use to illustrate the differences between different flavors of econometric practice—this has nothing to do with economic development but it is simple and easy to contrast with the development practice that I wish to discuss. In contrast to the models that I will discuss later, I think of this as a model in the spirit of the Cowles Foundation. It is the simplest possible Keynesian macroeconomic model of national income determination taken from once-standard econometrics textbooks. There are two equations that together comprise a complete macroeconomic system. The first equation is a consumption function in which aggregate consumption is a linear function of aggregate national income, while the second is the national income accounting identity that says that income is the sum of consumption and investment. I write the system in standard notation as

$$(1) \qquad C = \alpha + \beta Y + u,$$

$$(2) \qquad Y \equiv C + I.$$

According to (1), consumers choose the level of aggregate consumption with reference to their income, while in (2) investment is set by the "animal spirits" of entrepreneurs in some way that is outside of the model. No modern macroeconomist would take this model seriously, though the simple consumption function is an ancestor of more satisfactory and complete modern formulations; in particular, we can think of it (or at least its descendents) as being derived from a coherent model of intertemporal choice. Similarly, modern versions would postulate some theory for what determines investment I—here it is simply taken as given and assumed to be orthogonal to the consumption disturbance u.

In this model, consumption and income are simultaneously determined so that, in particular, a stochastic realization of u—consumers displaying animal spirits of their own—will affect not only C but also Y through equation (2), so that there is a positive correlation between u and Y. As a result, ordinary least squares (OLS) estimation of (1) will lead to upwardly biased and inconsistent estimates of the parameter β.

This simultaneity problem can be dealt with in a number of ways. One is to solve (1) and (2) to get the reduced form equations

$$(3) \quad C = \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta} I + \frac{u}{1-\beta},$$

$$(4) \quad Y = \frac{\alpha}{1-\beta} + \frac{I}{1-\beta} + \frac{u}{1-\beta}.$$

Both of these equations can be consistently estimated by OLS, and it is easy to show that the same estimates of α and β will be obtained from either one. An alternative method of estimation is to focus on the consumption function (1) and to use our knowledge of (2) to note that investment can be used as an instrumental variable (IV) for income. In the IV regression, there is a "first stage" regression in which income is regressed on investment; this is identical to equation (4), which is part of the reduced form. In the second stage, consumption is regressed on the predicted value of income from (4). In this simple case, the IV estimate of β is identical to the estimate from the reduced form. This simple model may not be a very good model—but it *is* a model, if only a primitive one.

I now leap forward sixty years and consider an apparently similar set up, again using an absurdly simple specification. The World Bank (let us imagine) is interested in whether to advise the government of China to build more railway stations as part of its poverty reduction strategy. The Bank economists write down an econometric model in which the poverty head count ratio in city c is taken to be a linear function of an indicator R of whether or not the city has a railway station,

$$(5) \qquad P_c = \gamma + \theta R_c + v_c,$$

where θ (I hesitate to call it a parameter) indicates the effect—presumably negative—of infrastructure (here a railway station) on poverty. While we cannot expect to get useful estimates of θ from OLS estimation of (5)—railway stations may be built to serve more prosperous cities, they are rarely built in deserts where there are no people, or there may be "third factors" that influence both—this is seen as a "technical problem" for which there is a wide range of econometric treatments including, of course, instrumental variables.

We no longer have the reduced form of the previous model to guide us but, if we can find an instrument Z that is correlated with whether a town has a railway station but uncorrelated with v, we can do the same calculations and obtain a consistent estimate. For the record, I write this equation

$$(6) \qquad R_c = \phi + \varphi Z_c + \eta_c.$$

Good candidates for Z might be indicators of whether the city has been designated by the Government of China as belonging to a special "infrastructure development area," or perhaps an earthquake that conveniently destroyed a selection of railway stations, or even the existence of river confluence near the city, since rivers were an early source of power, and railways served the power-based industries. I am making fun, but not much. And these instruments all have the real merit that there is some mechanism linking them to whether or not the town has a railway station, something that is not automatically guaranteed by the instrument being correlated with R and uncorrelated with v (see, for example, Peter C. Reiss and Frank A. Wolak 2007, pp. 4296–98).

My main argument is that the two econometric structures, in spite of their resemblance and the fact that IV techniques can be used for both, are in fact quite different. In particular, the IV procedures that work for the effect of national income on consumption are unlikely to give useful results for the effect of railway stations on poverty. To explain the differences, I begin with the language. In the original example, the reduced form is a fully specified system since it is derived from a notionally complete model of the determination of income. Consumption and income are treated symmetrically and appear as such in the reduced form equations (3) and (4). In contemporary examples, such as the railways, there is no complete theoretical system and there is no symmetry. Instead, we have a "main" equation (5), which used to be the "structural" equation (1). We also have a "first-stage" equation, which is the regression of railway stations on the instrument. The now rarely considered regression of the variable of interest on the instrument, here of poverty on earthquakes or on river confluences, is nowadays referred to as *the* reduced form, although it was originally one equation of a multiple equation reduced form—equation (6) is also part of the reduced form—within which it had no special significance. These language shifts sometimes cause confusion but they are not the most important differences between the two systems.

The crucial difference is that the relationship between railways and poverty is not a model at all, unlike the consumption model that embodied a(n admittedly crude) theory of income determination. While it is clearly *possible* that the construction of a railway station will reduce poverty, there are many possible mechanisms, some of which will work in one context and not in another. In consequence, θ is unlikely to be constant over different cities, nor can its variation be usefully thought of as random variation that is uncorrelated with anything else of interest. Instead, it is precisely the variation in θ that encapsulates the poverty reduction mechanisms that ought to be the main objects of our enquiry. Instead, the equation of interest—the so-called "main equation" (5)—is thought of as a representation of something more akin to an experiment or a biomedical trial in which some cities get "treated" with a station and some do not. The role of econometric analysis is not, as in the Cowles example, to estimate and investigate a casual model, but "to create an analogy, perhaps forced, between an observational study and an experiment" (David A. Freedman 2006, p. 691).

One immediate task is to recognize and somehow deal with the variation in θ, which is typically referred to as the "heterogeneity problem" in the literature. The obvious way is to define a parameter of interest in a way that corresponds to something we want to know for policy evaluation—perhaps the average effect on poverty over some group of cities—and then devise an appropriate estimation strategy. However, this step is often skipped in practice, perhaps because of a mistaken belief that the "main equation" (5) is a structural equation in which θ is a constant, so that the analysis can go immediately to the choice of instrument Z, over which a great deal of imagination and ingenuity is often exercised. Such ingenuity is often needed because it is difficult simultaneously to satisfy both of the standard criteria required for an instrument, that it be correlated with R_c and uncorrelated with v_c. However, if heterogeneity is indeed present, even satisfying the standard criteria is not sufficient to prevent the probability limit of the IV estimator depending on the choice of instrument (Heckman 1997). Without explicit prior consideration of the effect of the instrument choice on the parameter being estimated, such a procedure is effectively the opposite of standard statistical practice in which a parameter of interest is defined first, followed by an estimator that delivers that parameter. Instead, we have a procedure in which the choice of the instrument, which is guided by criteria designed for a situation in which there is no heterogeneity, is implicitly allowed to determine the parameter of interest. This goes beyond the old story of looking for an object where the light is strong enough to see; rather, we have at least some control over the light but choose to let it fall where it may and then proclaim that whatever it illuminates is what we were looking for all along.

Recent econometric analysis has given us a more precise characterization of what we can expect from such a method. In the railway example, where the instrument is the designation of a city as belonging to the "special infrastructure zone," the probability limit of the IV estimator is the average of poverty reduction effects over those cities who were induced to construct a railway station by being so designated. This average is known as the "local average treatment effect" (LATE) and its recovery by IV estimation requires a number of nontrivial conditions, including, for example, that no cities who would have constructed a railway station are perverse enough to be actually deterred from doing so by the positive designation (see Guido W. Imbens and Joshua D. Angrist 1994, who established the

LATE theorem). The LATE may or may not be a parameter of interest to the World Bank or the Chinese government and, in general, there is no reason to suppose that it will be. For example, the parameter estimated will typically *not* be the average poverty reduction effect over the designated cities, nor will it be the average effect over all cities.

I find it hard to make any sense of the LATE. We are unlikely to learn much about the processes at work if we refuse to say *anything* about what determines θ; heterogeneity is not a technical problem calling for an econometric solution but a reflection of the fact that we have not started on our proper business, which is trying to understand what is going on. Of course, if we are as skeptical of the ability of economic theory to deliver useful models as are many applied economists today, the ability to avoid modeling can be seen as an advantage, though it should not be a surprise when such an approach delivers answers that are hard to interpret. Note that my complaint is not with the "local" nature of the LATE—that property is shared by many estimation strategies and I will discuss later how we might overcome it. The issue here is rather the "average" and the lack of an ex ante characterization of the set over which the averaging is done. Angrist and Jörn-Steffen Pischke (2010) have recently claimed that the explosion of instrumental variables methods, including LATE estimation, has led to greater "credibility" in applied econometrics. I am not entirely certain what credibility means, but it is surely undermined if the parameter being estimated is not what we want to know. While in many cases what is estimated may be close to, or may contain information about, the parameter of interest, that this is actually so requires demonstration and is not true in general (see Heckman and Sergio Urzua 2009, who analyze cases where the LATE is an uninteresting and potentially misleading assemblage of parts of the underlying structure).

There is a related issue that bedevils a good deal of contemporary applied work, which is the understanding of *exogeneity*, a word that I have so far avoided. Suppose, for the moment, that the effect of railway stations on poverty is the same in all cities and we are looking for an instrument, which is required to be exogenous in order to consistently estimate θ. According to Merriam-Webster's dictionary, "exogenous" means "caused by factors or an agent from outside the organism or system," and this common usage is often employed in applied work. However, the consistency of IV estimation requires that the instrument be orthogonal to the error term v in the equation of interest, which is not implied by the Merriam-Webster definition (see Edward E. Leamer 1985, p. 260). Jeffrey M. Wooldridge (2002, p. 50) warns his readers that "you should not rely too much on the meaning of 'endogenous' from other branches of economics" and goes on to note that "the usage in econometrics, while related to traditional definitions, is used broadly to describe any situation where an explanatory variable is correlated with the disturbance." Heckman (2000) suggests using the term "external" (which he traces back to Wright and Frisch in the 1930s) for the Merriam-Webster definition, for variables whose values are not set or caused by the variables in the model and keeping "exogenous" for the orthogonality condition that is required for consistent estimation in this instrumental variable context. The terms are hardly standard, but I adopt them here because I need to make the distinction. The main issue, however, is not the terminology but that the two concepts be kept distinct so that we can see when the argument being *offered* is a justification for externality when what is *required* is a justification for exogeneity. An instrument that is external, but not exogeneous, will not yield consistent estimates of the parameter of interest, even when the parameter of interest is a constant.

An alternative approach is to keep the Miriam-Webster (or "other branches of economics") definition for exogenous and to require that, in addition to being exogenous, an instrument satisfy the "exclusion restrictions" of being uncorrelated with the disturbance. I have no objection to this usage, though the need to defend these additional restrictions is not always appreciated in practice. Yet exogeneity in this sense has no consequences for the consistency of econometric estimators and so is effectively meaningless.

Failure to separate externality and exogeneity—or to build a case for the validity of the exclusion restrictions—has caused, and continues to cause, endless confusion in the applied development (and other) literatures. Natural or geographic variables—distance from the equator (as an instrument for per capita GDP in explaining religiosity, Rachel M. McCleary and Robert J. Barro 2006), rivers (as an instrument for the number of school districts in explaining educational outcomes, Caroline M. Hoxby 2000), land gradient (as an instrument for dam construction in explaining poverty, Duflo and Rohini Pande 2007), or rainfall (as an instrument for economic growth in explaining civil war, Edward Miguel, Shanker Satyanath, and Ernest Sergenti 2004 and the examples could be multiplied ad infinitum)—are not affected by the variables being explained, and are clearly external. So are historical variables—the mortality of colonial settlers is not influenced by current institutional arrangements in ex-colonial countries (Daron Acemoglu, Simon Johnson, and James A. Robinson 2001) nor does the country's growth rate today influence the identity of their past colonizers (Barro 1998). Whether any of these instruments is *exogenous* (or satisfies the exclusion restrictions) depends on the specification of the equation of interest, and is not guaranteed by its *externality*. And because exogeneity is an identifying assumption that must be made prior to analysis of the data, empirical tests cannot settle the question. This does not prevent many attempts in the literature, often by misinterpreting a satisfactory *over*identification test as evidence for valid identification. Such tests can tell us whether estimates change when we select different subsets from a set of possible instruments. While the test is clearly useful and informative, acceptance is consistent with all of the instruments being invalid, while failure is consistent with a subset being correct. Passing an overidentification test does not validate instrumentation.

In my running example, earthquakes and rivers are external to the system and are neither caused by poverty nor by the construction of railway stations, and the designation as an infrastructure zone may also be determined by factors independent of poverty or railways. But even earthquakes (or rivers) are not exogenous if they have an effect on poverty other than through their destruction (or encouragement) of railway stations, as will almost always be the case. The absence of simultaneity does not guarantee exogeneity—exogeneity requires the absence of simultaneity but is not implied by it. Even random numbers—the ultimate external variables—may be endogenous, at least in the presence of heterogeneous effects if agents choose to accept or reject their assignment in a way that is correlated with the heterogeneity. Again, the example comes from Heckman's (1997) discussion of Angrist's (1990) famous use of draft lottery numbers as an instrumental variable in his analysis of the subsequent earnings of Vietnam veterans.

I can illustrate Heckman's argument using the Chinese railways example with the zone designation as instrument. Rewrite the equation of interest, (5), as

$$(7) \qquad P_c = \gamma + \bar{\theta} R_c + w_c$$

$$= \gamma + \bar{\theta} R_c + \{v_c + (\theta - \bar{\theta})\, R_c\},$$

where w_c is defined by the term in curly brackets, and $\bar{\theta}$ is the mean of θ over the cities that get the station so that the compound error term w has mean zero. Suppose the designation as an infrastructure zone is D_c, which takes values 1 or 0, and that the Chinese bureaucracy, persuaded by young development economists, decides to randomize and designates cities by flipping a yuan. For consistent estimation of $\bar{\theta}$, we want the covariance of the instrument with the error to be zero. The covariance is

$$(8)\quad E(D_c w_c) = E[(\theta - \bar{\theta})\,RD] = E[(\theta - \bar{\theta})\,|\,D = 1, R = 1] \times P(D = 1, R = 1),$$

which will be zero if either (*a*) the average effect of building a railway station on poverty among the cities induced to build one by the designation is the same as the average effect among those who would have built one anyway, or (*b*) no city not designated builds a railway station. If (*b*) is not guaranteed by *fiat*, we cannot suppose that it will otherwise hold, and we might reasonably hope that among the cities who build railway stations, those induced to do so by the designation are those where there is the largest effect on poverty, which violates (*a*). In the example of the Vietnam veterans, the instrument (the draft lottery number) fails to be exogenous because the error term in the earnings equation depends on each individual's rate of return to schooling, and whether or not each potential draftee accepted their assignment—their veteran's status—depends on that rate of return. This failure of exogeneity is referred to by Richard Blundell and Monica Costa Dias (2009) as selection on idiosyncratic gain and it adds to any bias caused by any failure of the instrument to be orthogonal to ν_c, ruled out here by the randomness of the instrument.

The general lesson is once again the ultimate futility of trying to avoid thinking about how and why things work—if we do not do so, we are left with undifferentiated heterogeneity that is likely to prevent consistent estimation of any parameter of interest. One appropriate response is to specify exactly how cities respond to their designation, an approach that leads to Heckman's local instrumental variable methods (Heckman and Edward Vytlacil 1999, 2007; Heckman, Urzua, and Vytlacil 2006). In a similar vein, David Card (1999) reviews estimates of the rate of return to schooling and explores how the choice of instruments leads to estimates that are averages over different subgroups of the population so that, by thinking about the implicit selection, evidence from different studies can be usefully summarized and compared. Similar questions are pursued in Gerard van den Berg (2008).

3. *Instruments of Development*

The question of whether aid has helped economies grow faster is typically asked within the framework of standard growth regressions. These regressions use data for many countries over a period of years, usually from the Penn World Table, the current version of which provides data on real per capita GDP and its components in purchasing power dollars for more than 150 countries as far back as 1950. The model to be estimated has the rate of growth of per capita GDP as the dependent variable, while the explanatory variables include the lagged value of GDP per capita, the share of investment in GDP, and measures of the educational level of the population (see, for example, Barro and Xavier Sala-i-Martin 1995, chapter 12, for an overview). Other variables are often added, and my main concern here is with one of these, external assistance (aid) as a fraction of GDP. A typical specification can be written

$$(9)\ \Delta\ln Y_{ct+1} = \beta_0 + \beta_1 \ln Y_{ct} + \beta_2 \frac{I_{ct}}{Y_{ct}} + \beta_3 H_{ct} + \beta_4 Z_{ct} + \theta A_{ct} + u_{ct},$$

where Y is per capita GDP, I is investment, H is a measure of human capital or education, and A is the variable of interest, aid as a share of GDP. Z stands for whatever other variables are included. The index c is for country and t for time. Growth is rarely measured on a year to year basis—the data in the Penn World Table are not suitable for annual analysis—so that growth may be measured over ten, twenty, or forty year intervals. With around forty years of data, there are four, two, or one observation for each country.

An immediate question is whether the growth equation (9) is a model-based Cowles-type equation, as in my national income example, or whether it is more akin to the atheoretical analysis in my invented Chinese railway example. There are elements of both here. If we ignore the Z and A variables in (9), the model can be thought of as a Solow growth model, extended to add human capital to physical capital (see again Barro and Sala-i-Martin, who derive their empirical specifications from the theory, and also N. Gregory Mankiw, David Romer, and David N. Weil 1992, who extended the Solow model to include education). However, the addition of the other variables, including aid, is typically less well justified. In some cases, for example under the assumption that all aid is invested, it is possible to calculate what effect we might expect aid to have (see Raghuram G. Rajan and Arvind Subramanian 2008). If we follow this route, (9) would not be useful—because aid is already included—and we should instead investigate *whether* aid is indeed invested, and then infer the effectiveness of aid from the effectiveness of investment. Even so, it presumably matters what kind of investment is promoted by aid, and aid for roads, for dams, for vaccination programs, or for humanitarian purposes after an earthquake are likely to have different effects on subsequent growth. More broadly, one of the main issues of contention in the debate is what aid actually does. Just to list a few of the possibilities, does aid increase investment, does aid crowd out domestic investment, is aid stolen, does aid create rent-seeking, or does aid undermine the institutions that are required for growth? Once all of these possibilities are admitted, it is clear that the analysis of (9) is not a Cowles model at all, but is seen as analogous to a biomedical experiment in which different countries are "dosed" with different amounts of aid, and we are trying to measure the average response. As in the Chinese railways case, a regression such as (9) will not give us what we want because the doses of aid are not randomly administered to different countries, so our first task is to find an instrumental variable that will generate quasi-randomness.

The most obvious problem with a regression of aid on growth is the simultaneous feedback from growth to aid that is generated by humanitarian responses to economic collapse or to natural or man-made disasters that engender economic collapse. More generally, aid flows from rich countries to poor countries, and poor countries, almost by definition, are those with poor records of economic growth. This feedback, from low growth to high aid, will obscure, nullify, or reverse any positive effects of aid. Most of the literature attempts to eliminate this feedback by using one or more instrumental variables and, although they would not express it in these terms, the aim of the instrumentation is to a restore a situation in which the pure effect of aid on growth can be observed as if in a randomized situation. How close we get to this ideal depends, of course, on the choice of instrument.

Although there is some variation across studies, there is a standard set of instruments, originally proposed by Peter Boone (1996), which includes the log of population size and various country dummies, for example, a dummy for Egypt or for francophone West Africa. One or both of these instruments are used in almost all the papers in a large subsequent literature, including Craig Burnside and David Dollar (2000), Henrik Hansen and Finn Tarp (2000, 2001), Carl-Johan Dalgaard and Hansen (2001), Patrick Guillaumont and Lisa Chauvet (2001), Robert Lensink and Howard White (2001), Easterly, Ross Levine, and David Roodman (2004), Dalgaard, Hansen, and Tarp (2004), Michael Clemens, Steven Radelet, and Rikhil Bhavnani (2004), Rajan and Subramanian (2008), and Roodman (2007). The rationale for population size is that larger countries get less aid per capita because the aid agencies allocate aid on a country basis, with less than full allowance for population size. The rationale for what I shall refer to as the "Egypt instrument" is that Egypt gets a great deal of American aid as part of the Camp David accords in which it agreed to a partial rapprochement with Israel. The same argument applies to the francophone countries, which receive additional aid from France because of their French colonial legacy. By comparing these countries with countries not so favored or by comparing populous with less populous countries, we can observe a kind of variation in the share of aid in GDP that is unaffected by the negative feedback from poor growth to compensatory aid. In effect, we are using the variation across populations of different sizes as a natural experiment to reveal the effects of aid.

If we examine the effects of aid on growth without any allowance for reverse causality, for example by estimating equation (9) by OLS, the estimated effect is typically negative. For example, Rajan and Subramanian (2008), in one of the most careful recent studies, find that an increase in aid by one percent of GDP comes with a reduction in the growth rate of one tenth of a percentage point a year. Easterly (2006) provides many other (sometimes spectacular) examples of negative associations between aid and growth. When instrumental variables are used to eliminate the reverse causality, Rajan and Subramanian find a weak or zero effect of aid and contrast that finding with the robust positive effects of investment on growth in specifications like (9). I should note that, although Rajan and Subramanian's study is an excellent one, it is certainly not without its problems and, as the authors note, there are many difficult econometric problems over and above the choice of instruments, including how to estimate dynamic models with country fixed effects on limited data, the choice of countries and sample period, the type of aid that needs to be considered, and so on. Indeed, it is those other issues that are the focus of most of the literature cited above. The substance of this debate is far from over.

My main concern here is with the use of the instruments, what they tell us, and what they might tell us. The first point is that neither the "Egypt" (or colonial heritage) nor the population instrument are plausibly exogenous; both are external—Camp David is not part of the model, nor was it caused by Egypt's economic growth, and similarly for population size—but exogeneity would require that neither "Egypt" nor population size have any influence on economic growth except through the effects on aid flows, which makes no sense at all. We also need to recognize the heterogeneity in the aid responses and try to think about how the different instruments are implicitly choosing different averages, involving different weightings or subgroups of countries. Or we could stop right here, conclude that there are no valid instruments, and that the aid to growth question is not answerable in this way. I shall argue otherwise, but I should also note that similar challenges over the validity of instruments have become routine in applied econometrics, leading to widespread skepticism by some, while others press on undaunted in an ever more creative search for exogeneity.

Yet consideration of the instruments is not without value, especially if we move away from instrumental variable estimation, with the use of instruments seen as technical, not substantive, and think about the reduced form which contains substantive information about the relationship between growth and the instruments. For the case of population size, we find that, conditional on the other variables, population size is unrelated to growth, which is one of the reasons that the IV estimates of the effects of aid are small or zero. This (partial) regression coefficient is a much simpler object than is the instrumental variable estimate; under standard assumptions, it tells us how much faster large countries grow than small countries, once the standard effects of the augmented Solow model have been taken into account. Does this tell us anything about the effectiveness of aid? Not directly, though it is surely useful to know that, while larger countries receive less per capita aid in relation to per capita income, they grow just as fast as countries that have received more, once we take into account the amount that they invest, their levels of education, and their starting level of GDP. But we would hardly conclude from this fact alone that aid does not increase growth. Perhaps aid works less well in small countries, or perhaps there is an offsetting positive effect of population size on economic growth. Both are possible and both are worth further investigation. More generally, such arguments are susceptible to fruitful discussions, not only among economists, but also with other social scientists and historians who study these questions, something that is typically difficult with instrumental variable methods. Economists' claims to methodological superiority based on instrumental variables ring particularly hollow when it is economists themselves who are so often misled. My argument is that, for both economists and noneconomists, the direct consideration of the reduced form is likely to generate productive lines of enquiry.

The case of the "Egypt" instrument is somewhat different. Once again the reduced form is useful (Egypt doesn't grow particularly fast in spite of all the aid it gets in consequence of Camp David), though mostly for making it immediately clear that the comparison of Egypt versus non-Egypt, or francophone versus nonfrancophone, is not a useful way of assessing the effectiveness of aid on growth. There is no reason to suppose that "being Egypt" has no effect on its growth other than through aid from the United States. Yet almost every paper in this literature unquestioningly uses the Egypt dummy as an instrument. Similar instruments based on colonial heritage face exactly the same problem; colonial heritage certainly affects aid, and colonial heritage is not influenced by current growth performance, but different colonists behaved differently and left different legacies of institutions and infrastructure, all of which have their own persistent effect on growth today.

The use of population size, an Egypt dummy, or colonial heritage variables as instruments in the analysis of aid effectiveness cannot be justified. These instruments are external, not exogenous, or if we use the Webster definition of exogeneity, they clearly fail the exclusion restrictions. Yet they continue in almost universal use in the aid-effectiveness literature, and are endorsed for this purpose by the leading exponents of IV methods (Angrist and Pischke 2010).

I conclude this section with an example that helps bridge the gap between analyses of the macro and analyses of the micro effects of aid. Many microeconomists agree that instrumentation in cross-country regressions is unlikely to be useful, while claiming that microeconomic analysis is capable of doing better. We may not be able to answer ill-posed questions about the macroeconomic effects of foreign assistance, but we can surely do better on specific projects and programs. Abhijit Vinayak Banerjee and Ruimin He (2008) have provided a list of the sort of studies that they like and that they believe should be replicated more widely. One of these, also endorsed by Duflo (2004), is a famous paper by Angrist and Victor Lavy (1999) on whether schoolchildren do better in smaller classes, a position frequently endorsed by parents and by teacher's unions but not always supported by empirical work. The question is an important one for development assistance because smaller class sizes cost more and are a potential use for foreign aid. Angrist and Lavy's paper uses a natural experiment, not a real one, and relies on IV estimation, so it provides a bridge between the relatively weak natural experiments in this section, and the actual randomized controlled trials in the next.

Angrist and Lavy's study is about the allocation of children enrolled in a school into classes. Many countries set their class sizes to conform to some version of Maimonides' rule, which sets a maximum class size, beyond which additional teachers must be found. In Israel, the maximum class size is set at 40. If there are less than 40 children enrolled, they will all be in the same class. If there are 41, there will be two classes, one of 20, and one of 21. If there are 81 or more children, the first two classes will be full, and more must be set up. Angrist and Lavy's figure 1 plots actual class size and Maimonides' rule class size against the number of children enrolled; this graph starts off running along the 45-degree line, and then falls discontinuously to 20 when enrollment is 40, increasing with slope of 0.5 to 80, falling to 27.7 (80 divided by 3) at 80, rising again with a slope of 0.25, and so on. They show that actual class sizes, while not exactly conforming to the rule, are strongly influenced by it and exhibit the same saw tooth pattern. They then plot test scores against enrollment, and show that they display the opposite pattern, rising at each of the discontinuities where class size abruptly falls. This is a natural experiment, with Maimonides' rule inducing quasi-experimental variation, and generating a predicted class size for each level of enrollment which serves as an instrumental variable in a regression of test scores on class size. These IV estimates, unlike the OLS estimates, show that children in smaller classes do better.

Angrist and Lavy's paper, the creativity of its method, and the clarity of its result has set the standard for micro empirical work since it was published, and it has had a far-reaching effect on subsequent empirical work in labor and development economics. Yet there is a problem, which has become apparent over time. Note first the heterogeneity; it is improbable that the effect of lower class size is the same for all children so that, under the assumptions of the LATE theorem, the IV estimate recovers a weighted average of the effects for those children who are shifted by Maimonides' rule from a larger to a smaller class. Those children might not be the same as other children, which makes it hard to know how useful the numbers might be in other contexts, for example when all children are put in smaller class sizes. The underlying reasons for this heterogeneity are not addressed in this quasi-experimental approach. To be sure of what is happening here, we need to know more about how different children finish up in different classes, which raises the possibility that the variation across the discontinuities may not be orthogonal to other factors that affect test scores.

A recent paper by Miguel Urquiola and Eric Verhoogen (2009) explores how it is that children are allocated to different class sizes in a related, but different, situation in Chile where a version of Maimonides' rule is in place. Urquiola and Verhoogen note that parents care a great deal about whether their children are in the 40 child class or the 20 child class, and for the private schools they study, they construct a model in which there is sorting across the boundary, so that the children in the smaller classes have richer, more educated parents than the children in the larger classes. Their data match such a model, so that at least some of the differences in test scores across class size come from differences in the children that would be present whatever the class size. This paper is an elegant example of why it is so dangerous to make inferences from natural experiments without understanding the mechanisms at work. It also strongly suggests that the question of the effects of class size on student performance is not well-defined without a description of the environment in which class size is being changed (see also Christopher A. Sims 2010).

Another good example comes to me in private correspondence from Branko Milanovic, who was a child in Belgrade in a school only half of whose teachers could teach in English, and who was randomly assigned to a class taught in Russian. He remembers losing friends whose parents (correctly) perceived the superior value of an English education, and were insufficiently dedicated socialists to accept their assignment. The two language groups of children remaining in the school, although "randomly" assigned, are far from identical, and IV estimates using the randomization could overstate the superiority of the English medium education.

More generally, these are examples where an instrument induces actual or quasi-random assignment, here of children into different classes, but where the assignment can be undone, at least partially, by the actions of the subjects. If children—or their parents—care about whether they are in small or large classes, or in Russian or English classes, some will take evasive action—by protesting to authorities, finding a different school, or even moving—and these actions will generally differ by rich and poor children, by children with more or less educated parents, or by any factor that affects the cost to the child of being in a larger class. The behavioral response to the quasi-randomization (or indeed randomization) means that the groups being compared are not identical to start with (see also Justin McCrary 2008 and David S. Lee and Thomas Lemieux 2009 for further discussion and for methods of detection when this is happening).

In preparation for the next section, I note that the problem here is *not* the fact that we have a quasi-experiment rather than a real experiment, so that there was no actual randomization. If children had been randomized into class size, as in the Belgrade example, the problems would have been the same unless there had been some mechanism for forcing the children (and their parents) to accept the assignment.

4. *Randomization in the Tropics*

Skepticism about econometrics, doubts about the usefulness of structural models in economics, and the endless wrangling over identification and instrumental variables has led to a search for alternative ways of learning about development. There has also been frustration with the World Bank's apparent failure to learn from its own projects and its inability to provide a convincing argument that its past activities have enhanced economic growth and poverty reduction. Past development practice is seen as a succession of fads, with one supposed magic bullet replacing another—from planning to infrastructure to human capital to structural adjustment to health and social capital to the environment and back to infrastructure—a process that seems not to be guided by progressive learning. For many economists, and particularly for the group at the Poverty Action Lab at MIT, the solution has been to move toward randomized controlled trials of projects, programs, and policies. RCTs are seen as generating gold standard evidence that is superior to econometric evidence and that is immune to the methodological criticisms that are typically directed at econometric analyses. Another aim of the program is to persuade the World Bank to replace its current evaluation methods with RCTs; Duflo (2004) argues that randomized trials of projects would generate knowledge that could be used elsewhere, an international public good. Banerjee (2007b, chapter 1) accuses the Bank of "lazy thinking," of a "resistance to knowledge," and notes that its recommendations for poverty reduction and empowerment show a striking "lack of distinction made between strategies founded on the hard evidence provided by randomized trials or natural experiments and the rest." In all this there is a close parallel with the evidence-based movement in medicine that preceded it, and the successes of RCTs in medicine are frequently cited. Yet the parallels are almost entirely rhetorical, and there is little or no reference to the dissenting literature, as surveyed for example in John Worrall (2007) who documents the rise and fall in medicine of the rhetoric used by Banerjee. Nor is there any recognition of the many problems of medical RCTs, some of which I shall discuss as I go.

The movement in favor of RCTs is currently very successful. The World Bank is now conducting substantial numbers of randomized trials, and the methodology is sometimes explicitly requested by governments, who supply the World Bank with funds for this purpose (see World Bank 2008 for details of the Spanish Trust Fund for Impact Evaluation). There is a new International Initiative for Impact Evaluation which "seeks to improve the lives of poor people in low- and middle-income countries by providing, and summarizing, evidence of what works, when, why and for how much," (International Initiative for Impact Evaluation 2008), although not exclusively by randomized controlled trials. The Poverty Action Lab lists dozens of completed and ongoing projects in a large number of countries, many of which are project evaluations. Many development economists would join many physicians in subscribing to the jingoist view proclaimed by the editors of the *British Medical Journal* (quoted by Worrall 2007, p. 996) which noted that "Britain has given the world Shakespeare, Newtonian physics, the theory of evolution, parliamentary democracy—and the randomized trial."

4.1 *The Ideal RCT*

Under ideal conditions and when correctly executed, an RCT can estimate certain quantities of interest with minimal assumptions, thus absolving RCTs of one complaint against econometric methods—that they rest on often implausible economic models. It is useful to lay out briefly the (standard) framework for these results, originally due to Jerzy Neyman in the 1920s, currently often referred to as the Holland–Rubin framework or the Rubin causal model (see Freedman 2006 for a discussion of the history). According to this, each member of the population under study, labeled i, has two possible values associated with it, Y_{0i} and Y_{1i}, which are the outcomes that i would display if it did not get the treatment, $T_i = 0$, and if it did get the treatment, $T_i = 1$. Since each i is either in the treatment group or in the control group, we observe one of Y_{0i} and Y_{1i} but not both. We would like to know something about the distribution over i of the effects of the treatment, $Y_{i1} - Y_{i0}$, in particular its mean $\overline{Y}_1 - \overline{Y}_0$. In a sense, the most surprising thing about this set-up is that we can say anything at all without further assumptions or without any modeling. But that is the magic that is wrought by the randomization.

What we *can* observe in the data is the difference between the average outcome in the treatments and the average outcome in the controls, or $E(Y_i \mid T_i = 1) - E(Y_i \mid T_i = 0)$. This difference can be broken up into two terms

$$(10)\quad E(Y_{i1} \mid T_i = 1) - E(Y_{i0} \mid T_i = 0)$$
$$= [E(Y_{i1} \mid T_i = 1) - E(Y_{i0} \mid T_i = 1)]$$
$$+ [E(Y_{i0} \mid T_i = 1) - E(Y_{i0} \mid T_i = 0)].$$

Note that on the right hand side the second term in the first square bracket cancels out with the first term in the second square bracket. But the term in the second square bracket is zero by randomization; the nontreatment outcomes, like any other characteristic, are identical in expectation in the control and treatment groups. We can therefore write (10) as

$$(11)\quad E(Y_{i1} \mid T_i = 1) - E(Y_{i0} \mid T_i = 0)$$
$$= [E(Y_{i1} \mid T_i = 1) - E(Y_{i0} \mid T_i = 1)],$$

so that the difference in the two observable outcomes is the difference between the average treated outcome and the average untreated outcome in the treatment group. The last term on the right hand side would be unobservable in the absence of randomization.

We are not quite done. What we would like is the average of the difference, rather than the difference of averages that is currently on the right-hand side of (11). But the expectation is a linear operator, so that the difference of the averages is identical to the average of the differences, so that we reach, finally

$$(12)\quad E(Y_{i1} \mid T_i = 1) - E(Y_{i0} \mid T_i = 0)$$
$$= [E(Y_{i1} - Y_{i0} \mid T_i = 1).$$

The difference in means between the treatments and controls is an estimate of the average treatment effect among the treated, which, since the treatment and controls differ only by randomization, is an estimate of the average treatment effect for all. This standard but remarkable result depends both on randomization and on the linearity of expectations.

One immediate consequence of this derivation is a fact that is often quoted by critics of RCTs, but often ignored by practitioners, at least in economics: RCTs are informative about the *mean* of the treatment effects, $Y_{i1} - Y_{i0}$, but do not identify other features of the distribution. For example, the median of the difference is not the difference in medians, so an RCT is not, by itself, informative about the median treatment effect, something that could be of as much interest to policymakers as the mean treatment effect. It might also be useful to know the fraction of the population for which the treatment effect is positive, which once again is not identified from a trial. Put differently, the trial might reveal an average positive effect although nearly all of the population is hurt with a few receiving very large benefits, a situation that cannot be revealed by the RCT, although it might be disastrous if implemented. Indeed, Ravi Kanbur (2001) has argued that much of the disagreement about development policy is driven by differences of this kind.

Given the minimal assumptions that go into an RCT, it is not surprising that it cannot tell us everything that we would like to know. Heckman and Smith (1995) discuss these issues at greater length and also note that, in some circumstances, more can be learned. Essentially, the RCT gives us two marginal distributions from which we would like to infer a joint distribution; this is impossible, but the marginal distributions limit the joint distribution in ways that can be useful. For example, Charles F. Manski (1996) notes that a planner who is maximizing the expected value of a social welfare function needs only the two marginal distributions to check the usefulness of the treatment. Beyond that, if the probability distribution of outcomes among the treated stochastically dominates the distribution among the controls, we know that appropriately defined classes of social welfare functions will show an improvement without having to know what the social welfare function is. Not all relevant cases are covered by these examples; even if a drug saves lives on average, we need to know whether it is uniformly beneficial or kills some and saves more. To answer such questions, we will have to make assumptions beyond those required for an RCT; as usual, some questions can be answered with fewer assumptions than others.

In practice, researchers who conduct randomized controlled trials often do present results on statistics other than the mean. For example, the results can be used to run a regression of the form

$$(13) \qquad Y_i = \beta_0 + \beta_1 T_i + \sum_j \theta_j X_{ij} + \sum_j \phi_j X_{ij} \times T_i + u_i,$$

where T is a binary variable that indicates treatment status, and the X's are various characteristics measured at baseline that are included in the regression both on their own (main effects) and as interactions with treatment status (see Suresh de Mel, David McKenzie, and Christopher Woodruff 2008 for an example of a field experiment with micro-enterprises in Sri Lanka). The estimated treatment effect now varies across the population, so that it is possible, for example, to estimate whether the average treatment effect is positive or negative for various subgroups of interest. These estimates depend on more assumptions than the trial itself, in particular on the validity of running a regression like (13), on which I shall have more to say below. One immediate charge against such a procedure is data mining. A sufficiently determined examination of any trial will eventually reveal some subgroup for which the treatment yielded a significant effect of some sort, and there is no general way of adjusting standard errors to protect against the possibility. A classic example from medicine comes from the ISIS-2 trial of the use of aspirin after heart attacks (Richard Peto, Rory Collins, and Richard Gray 1995). A randomized trial established a beneficial effect with a significance level of better than 10^{-6}, yet ex post analysis of the data showed that there was no significant effect for trial subjects whose astrological signs were Libra or Gemini. In drug trials, the FDA rules require that analytical plans be submitted prior to trials and drugs cannot be approved based on ex post data analysis. As noted by Robert J. Sampson (2008), one analysis of the recent Moving to Opportunity experiment has an appendix listing tests of many thousands of outcomes (Lisa Sanbonmatsu et al. 2006).

I am not arguing against posttrial subgroup analysis, only that, as is enshrined in the FDA rules, any special epistemic status (as in "gold standard," "hard," or "rigorous" evidence) possessed by RCTs does not extend to ex post subgroup analysis if only because there is no guarantee that a new RCT on post-experimentally defined subgroups will yield the same result. Such analyses do not share any special evidential status that might arguably be accorded to RCTs and must be assessed in exactly the same way as we would assess any nonexperimental or econometric study. These issues are wonderfully exposed by the subgroup analysis of drug effectiveness by Ralph I. Horwitz et al. (1996), criticized by Douglas G. Altman (1998), who refers to such studies as "a false trail," by Stephen Senn and Frank Harrell (1997), who call them "wisdom after the event," and by George Davey Smith and Matthias Egger (1998), who call them "incommunicable knowledge," drawing the response by Horwitz et al. (1997) that their critics have reached "the tunnel at the end of the light." While it is clearly absurd to discard data because we do not know how to analyze it with sufficient purity and while many important findings have come from posttrial analysis of experimental data, both in medicine and in economics, for example of the negative income tax experiments of the 1960s, the concern about data-mining remains real enough. In large-scale, expensive trials, a zero or very small result is unlikely to be welcomed, and there is likely to be considerable pressure to search for some subpopulation or some outcome that shows a more palatable result, if only to help justify the cost of the trial.

The mean treatment effect from an RCT may be of limited value for a physician or a policymaker contemplating specific patients or policies. A new drug might do better than a placebo in an RCT, yet a physician might be entirely correct in not prescribing it for a patient whose characteristics, according to the physician's theory of the disease, might lead her to suppose that the drug would be harmful. Similarly, if we are convinced that dams in India do not reduce poverty on average, as in Duflo and Pande's (2007) IV study, there is no implication about any specific dam, even one of the dams included in the study, yet it is always a specific dam that a policymaker has to approve. Their evidence certainly puts a higher burden of proof on those proposing a new dam, as would be the case for a physician prescribing in the face of an RCT, though the force of the evidence depends on the size of the mean effect and the extent of the heterogeneity in the responses. As was the case with the material discussed in sections 2 and 3, heterogeneity poses problems for the analysis of RCTs, just as it posed problems for nonexperimental methods that sought to approximate randomization. For this reason, in his *Planning of Experiments*, David R. Cox (1958, p. 15) begins his book with the *assumption* that the treatment effects are identical for all subjects. He notes that the RCT will still estimate the mean treatment effect with heterogeneity but argues that such estimates are "quite misleading," citing the example of two internally homogeneous subgroups with distinct average treatment effects, so that the RCT delivers an estimate that applies to no one. Cox's recommendation makes a good deal of sense when the experiment is being applied to the parameter of a well-specified model, but it could not be further away from most current practice in either medicine or economics.

One of the reasons why subgroup analysis is so hard to resist is that researchers, however much they may wish to escape the straitjacket of theory, inevitably have some mechanism in mind, and some of those mechanisms can be "tested" on the data from the trial. Such "testing," of course, does not satisfy the strict evidential standards that the RCT has been set up to satisfy and, if the investigation is constrained to satisfy those standards, no ex post speculation is permitted. Without a prior theory and within its own evidentiary standards, an RCT targeted at "finding out what works" is not informative about mechanisms, if only because there are always multiple mechanisms at work. For example, when two independent but identical RCTs in two cities in India find that children's scores improved less in Mumbai than in Vadodora, the authors state "this is likely related to the fact that over 80 percent of the children in Mumbai had already mastered the basic language skills the program was covering" (Duflo, Rachel Glennerster, and Michael Kremer 2008). It is not clear how "likely" is established here, and there is certainly no evidence that conforms to the "gold standard" that is seen as one of the central justifications for the RCTs. For the same reason, repeated *successful* replications of a "what works" experiment, i.e., one that is unrelated to some underlying or guiding mechanism, is both unlikely and unlikely to be persuasive. Learning about theory, or mechanisms, requires that the investigation be targeted toward that theory, toward *why* something works, not *whether* it works. Projects can rarely be replicated, though the mechanisms underlying success or failure will often be replicable and transportable. This means that, if the World Bank had indeed randomized all of its past projects, it is unlikely that the cumulated evidence would contain the key to economic development.

Cartwright (2007a) summarizes the benefits of RCTs relative to other forms of evidence. In the ideal case, "if the assumptions of the test are met, a positive result *implies* the appropriate causal conclusion," that the intervention "worked" and caused a positive outcome. She adds "the benefit that the conclusions follow deductively in the ideal case comes with great cost: narrowness of scope" (p. 11).

4.2 *Tropical RCTs in Practice*

How well do actual RCTs approximate the ideal? Are the assumptions generally met in practice? Is the narrowness of scope a price that brings real benefits or is the superiority of RCTs largely rhetorical? RCTs allow the investigator to induce variation that might not arise nonexperimentally, and this variation can reveal responses that could never have been found otherwise. Are these responses the relevant ones? As always, there is no substitute for examining each study in detail, and there is certainly nothing in the RCT methodology itself that grants immunity from problems of implementation. Yet there are some general points that are worth discussion.

The first is the seemingly obvious practical matter of how to compute the results of a trial. In theory, this is straightforward—we simply compare the mean outcome in the experimental group with the mean outcome in the control group and the difference is the causal effect of the intervention. This simplicity, compared with the often baroque complexity of econometric estimators, is seen as one of the great advantages of RCTs, both in generating convincing results and in explaining those results to policymakers and the lay public. Yet any difference is not useful without a standard error and the calculation of the standard error is rarely quite so straightforward. As Ronald A. Fisher (1935) emphasized from the very beginning, in his famous discussion of the tea lady, randomization plays two separate roles. The first is to guarantee that the probability law governing the selection of the control group is the same as the probability law governing the selection of the experimental group. The second is to provide a probability law that enables us to judge whether a difference between the two groups is significant. In his tea lady example, Fisher uses combinatoric analysis to calculate the exact probabilities of each possible outcome, but in practice this is rarely done.

Duflo, Glennerster, and Kremer (2008, p. 3921) (DGK) explicitly recommend what seems to have become the standard method in the development literature, which is to run a restricted version of the regression (13), including only the constant and the treatment dummy,

$$(14) \qquad Y_i = \beta_0 + \beta_1 T_i + u_i.$$

As is easily shown, the OLS estimate of β_1 is simply the difference between the mean of the Y_i in the experimental and control groups, which is exactly what we want. However, the *standard error* of β_1 from the OLS regression is not generally correct. One problem is that the variance among the experimentals may be different from the variance among the controls, and to assume that the experiment *does not* affect the variance is very much against the minimalist spirit of RCTs. If the regression (14) is run with the standard heteroskedasticity correction to the standard error, the result will be the same as the formula for the standard error of the difference between two means, but not otherwise except in the special case where there are equal numbers of experimental and controls, in which case it turns out that the correction makes no difference and the OLS standard error is correct. It is not clear in the experimental development literature whether the correction is routinely done in practice, and the handbook review by DGK makes no mention of it, although it provides a thoroughly useful review of many other aspects of standard errors.

Even with the correction for unequal variances, we are not quite done. The general problem of testing the significance of the differences between the means of two normal populations with different variances is known as the Fisher–Behrens problem. The test statistic computed by dividing the difference in means by its estimated standard error does not have the t–distribution when the variances are different in treatments and controls, and the significance of the estimated difference in means is likely to be overstated if no correction is made. If there are equal numbers of treatments and controls, the statistic will be approximately distributed as Student's t but with degrees of freedom that can be as little as half the nominal degrees of freedom when one of the two variances is zero. In general, there is also no reason to suppose that the heterogeneity in the treatment effects is normal, which will further complicate inference in small samples.

Another standard practice, recommended by DGK, and which is also common in medical RCTs according to Freedman (2008), is to run the regression (14) with additional controls taken from the baseline data or equivalently (13) with the X_i but without the interactions,

$$(15)\; Y_i = \beta_0 + \beta_1 T_i + \sum_j \theta_j X_{ij} + u_i.$$

The standard argument is that, if the randomization is done correctly, the X_i will be orthogonal to the treatment variable T_i so that their inclusion does not affect the estimate of β_1, which is the parameter of interest. However, by absorbing variance, as compared with (14), they will increase the precision of the estimate—this is not necessarily the case, but will often be true. DGK (p. 3924) give an example: "controlling for baseline test scores in evaluations of educational interventions greatly improves the precision of the estimates, which reduces the cost of these evaluations when a baseline test can be conducted."

There are two problems with this procedure. The first, which is noted by DGK, is that, as with posttrial subgroup analysis, there is a risk of data mining—trying different control variables until the experiment "works"—unless the control variables are specified in advance. Again, it is hard to tell whether or how often this dictum is observed. The second problem is analyzed by Freedman (2008), who notes that (15) is not a standard regression because of the heterogeneity of the responses. Write α_i for the (hypothetical) treatment response of unit i, so that, in line with the discussion in the previous subsection, $\alpha_i = Y_{i1} - Y_{i0}$, and we can write the identity

$$(16)\qquad \begin{aligned} Y_i &= Y_{i0} + \alpha_i T_i \\ &= \overline{Y}_0 + \alpha_i T_i + (Y_{i0} - \overline{Y}_0), \end{aligned}$$

which looks like the regression (15) with the X's and the error term capturing the variation in Y_{i0}. The only difference is that the coefficient on the treatment term has an i suffix because of the heterogeneity. If we define $\alpha = E(\alpha_i \mid T_i = 1)$, the average treatment effect among the treated, as the parameter of interest, as in section 4.1, we can rewrite (16) as

$$(17) \quad Y_i = \overline{Y}_0 + \alpha T_i + (Y_{i0} - \overline{Y}_0) + (\alpha_i - \alpha) T_i.$$

Finally, and to illustrate, suppose that we model the variation in Y_{i0} as a linear function of an observable scalar X_{i0} and a residual η_i, we have

$$(18) \quad Y_i = \beta_0 + \alpha T_i + \theta(X_i - \overline{X}) + [\eta_i + (\alpha_i - \alpha) T_i],$$

with $\beta_0 = Y_0$, which is in the regression form (15) but allows us to see the links with the experimental quantities.

Equation (18) is analyzed in some detail by Freedman (2008). It is easily shown that T_i is orthogonal to the compound error, but that this is not true of $X_i - \overline{X}$. However, the two right hand side variables are uncorrelated because of the randomization, so the OLS estimate of $\beta_1 = \alpha$ is consistent. This is not true of θ, though this may not be a problem if the aim is simply to reduce the sampling variance. A more serious issue is that the dependency between T_i and the compound error term means that the OLS estimate of the average treatment effect α is biased, and in small samples this bias—which comes from the heterogeneity—may be substantial. Freedman notes that the leading term in the bias of the estimate of the OLS estimate of α is φ/n where n is the sample size and

$$(19) \quad \varphi = -\lim \frac{1}{n} \sum_{i=1}^{n} (\alpha_i - \alpha) Z_i^2,$$

where Z_i is the standardized (z-score) version of X_i. Equation (19) shows that the bias comes from the heterogeneity, or more specifically, from a covariance between the heterogeneity in the treatment effects and the squares of the included covariates. With the sample sizes typically encountered in these experiments, which are often expensive to conduct, the bias can be substantial. One possible strategy here would be to compare the estimates of α with and without covariates; even ignoring pretest bias, it is not clear how to make such a comparison without a good estimate of the standard error. Alternatively, and as noted by Imbens (2009) in his commentary on this paper, it is possible to remove bias using a "saturated" regression model, for example by estimating (15) when the covariates are discrete and there is a complete set of interactions. This is equivalent to stratifying on each combination of values of the covariates, which diminishes any effect on reducing the standard errors and, unless the stratification is done ex ante, raises the usual concerns about data mining.

Of these and related issues in medical trials, Freedman (2008, p. 13) writes "Practitioners will doubtless be heard to object that they know all this perfectly well. Perhaps, but then why do they so often fit models without discussing assumptions?"

All of the issues so far can be dealt with, either by appropriately calculating standard errors or by refraining from the use of covariates, though this might involve drawing larger and more expensive samples. However, there are other practical problems that are harder to fix. One of these is that subjects may fail to accept assignment, so that people who are assigned to the experimental group may refuse, and controls may find a way of getting the treatment, and either may drop out of the experiment altogether. The classical remedy of double blinding, so that neither the subject nor the experimenter know which subject is in which group, is rarely feasible in social experiments—children know their class size—and is often not feasible in medical trials—subjects may decipher the randomization, for example by asking a laboratory to check that their medicine is not a placebo. Heckman (1992) notes that, in contrast to people, "plots of grounds do not respond to anticipated treatments of fertilizer, nor can they excuse themselves from being treated." This makes the important point, further developed by Heckman in later work, that the deviations from assignment are almost certainly purposeful, at least in part. The people who struggle to escape their assignment will do so more vigorously the higher are the stakes, so that the deviations from assignment cannot be treated as random measurement error, but will compromise the results in fundamental ways.

Once again, there is a widely used technical fix, which is to run regressions like (15) or (18), with *actual* treatment status in place of the *assigned* treatment status T_i. This replacement will destroy the orthogonality between treatment and the error term, so that OLS estimation will no longer yield a consistent estimate of the average treatment effect among the treated. However, the assigned treatment status, which is known to the experimenter, is orthogonal to the error term and is correlated with the actual treatment status, and so can serve as an instrumental variable for the latter. But now we are back to the discussion of instrumental variables in section 2, and we are doing econometrics, not an ideal RCT. Under the assumption of no "defiers"—people who do the opposite of their assignment just because of the assignment (and it is not clear "just why are there no defiers" Freedman 2006)—the instrumental variable converges to the LATE. As before, it is unclear whether this is what we want, and there is no way to find out without modeling the behavior that is responsible for the heterogeneity of the response to assignment, as in the local instrumental variable approach developed by Heckman and his coauthors, Heckman and Vytlacil (1999, 2007). Alternatively, and as recommended by Freedman (2005, p. 4; 2006), it is always informative to make a simple unadjusted comparison of the average outcomes between treatments and controls according to the original assignment. This may also be enough if what we are concerned with is whether the treatment works or not, rather than with the size of the effect. In terms of instrumental variables, this is a recommendation to look at the reduced form, and again harks back to similar arguments in section 2 on aid effectiveness.

One common problem is that the people who agree to participate in the experiment (as either experimental or controls) are not themselves randomly drawn from the general population so that, even if the experiment itself if perfectly executed, the results are not transferable from the experimental to the parent population and will not be a reliable guide to policy in the parent population. In effect, the selection or omitted variable bias that is a potential problem in nonexperimental studies comes back in a different form and, without an analysis of the two biases, it is impossible to conclude which estimate is better—a biased nonexperimental analysis might do better than a randomized controlled trial if enrollment into the trial is nonrepresentative. In drug trials, ethical protocols typically require the principle of "equipoise"—that the physician, or at least physicians in general, believe that the patient has an equal chance of improvement with the experimental and control drug. Yet risk-averse patients will not accept such a gamble, so that there is selection into treatment, and the requirements of ethics and of representativity come into direct conflict. While this argument does not apply to all trials, there are many ways in which the experimental (experimentals plus controls) and parent populations can differ.

There are also operational problems that afflict every actual experiment; these can be mitigated by careful planning—in RCTs compared with econometric analysis, most of the work is done before data collection, not after—but not always eliminated.

In this context, I turn to the flagship study of the new movement in development economics—Miguel and Kremer's (2004) study of intestinal worms in Kenya. This paper is repeatedly cited in DGK's manual and it is one of the exemplary studies cited by Duflo (2004) and by Banerjee and He (2008). It was written by two senior authors at leading research universities and published in the most prestigious technical journal in economics. It has also received a great deal of positive attention in the popular press (see, for example, David Leonhardt 2008) and has been influential in policy (see Poverty Action Lab 2007). In this study, a group of "seventy-five rural primary schools were phased into treatment in a randomized order," with the finding "that the program reduced school absenteeism by at least one quarter, with particularly large participation gains among the youngest children, making deworming a highly effective way to boost school participation among young children" (p. 159). The point of the RCT is less to show that deworming medicines are effective, but to show that school-based treatment is more effective than individual treatment because children infect one another. As befits a paper that aims to change method, there is emphasis on the virtues of randomization, and the word "random" or its derivatives appears some sixty times in the paper. However, the "randomization" in the study is actually an assignment of schools to three groups by their order in the alphabet as in Albert Infant School to group 1, Alfred Elementary to group 2, Bell's Academy to group 3, Christopher School to group 1 again, Dean's Infants to group 2, and so on. Alphabetization, not randomization, was also used in the experiment on flip charts in schools by Paul Glewwe et al. (2004); this paper, like "Worms," is much cited as evidence in favor of the virtues of randomization.

Alphabetization may be a reasonable solution when randomization is impossible but we are then in the world of quasi- or natural experiments, not randomized experiments; in the latter, the balance of observable and unobservable factors in treatments and controls is guaranteed by design, at least in expectation, in the former, it has to be argued for each case, and the need for such argument is one of the main claims for the superiority of the randomized approach. As is true with all forms of quasi-randomization, alphabetization does not guarantee orthogonality with potential confounders, however plausible it may be in the Kenyan case. Resources are often allocated alphabetically because that is how many lists are presented (see, for example, Štěpán Jurajda and Daniel Münich 2009 for documentation of students being admitted into selective schools (partly) based on their position in the alphabet). If this were the case in Kenya, schools higher in the alphabet would be systematically different and this difference would be inherited in an attenuated form by the three groups. Indeed, this sort of contamination is described by Cox (1958, pp. 74–75) who explicitly warns against such designs. Of course, it is also possible that, in this case, the alphabetization causes no confounding with factors known or unknown. If so, there is still an issue with the calculation of standard errors. Without a probability law, we have no way of discovering whether the difference between treatments and controls could have arisen by chance. We might think of modeling the situation here by imagining that the assignment was equivalent to taking a random starting value and assigning every third school to treatment. If so, the fact that there are only three possible assignments of schools would have to be taken into account in calculating the standard errors, and nothing of this kind is reported. As it is, it is impossible to tell whether the experimental differences in these studies are or are not due to chance.

In this subsection, I have dwelt on practice not to critique particular studies or particular results; indeed it seems entirely plausible that deworming is a good idea and that the costs are low relative to other interventions. My main point here is different—that conducting good RCTs is exacting and often expensive, so that problems often arise that need to be dealt with by various econometric or statistical fixes. There is nothing wrong with such fixes in principle—though they often compromise the substance, as in the instrumental variable estimation to correct for failure of assignment—but their application takes us out of the world of ideal RCTs and back into the world of everyday econometrics and statistics. So that RCTs, although frequently useful, carry no special exemption from the routine statistical and substantive scrutiny that should be routinely applied to any empirical investigation.

Although it is well beyond my scope in this paper, I should note that RCTs in medicine—the gold standard to which development RCTs often compare themselves—also encounter practical difficulties, and their primacy is not without challenge. In particular, ethical (human subjects) questions surrounding RCTs in medicine have become severe enough to seriously limit what can be undertaken, and there is still no general agreement on a satisfactory ethical basis for RCTs. Selection into medical trials is not random from the population; in particular, patients are typically excluded if they suffer from conditions other than those targeted in the trial, so that researchers can examine effects without contamination from comorbidities. Yet many actual patients do have comorbidities—this is particularly true among the elderly—so drugs are frequently prescribed to those who were not represented in the original trials (Jerome Groopman 2009). In RCTs of some medical procedures, the hospitals chosen to participate are carefully selected, and favorable trial results may not be obtainable elsewhere (see David E. Wennberg et al. 1998 for an example in which actual mortality is many times higher than in the trials, sufficiently so as to reverse the desirability of the adoption of the procedure). There is also a concern that those who sponsor trials, those who analyze them, and those who set evidence-based guidelines using them sometimes have financial stakes in the outcome, which can cast doubts on the results. This is currently not a problem in economics but would surely become one if, as the advocates argue, successful RCTs became a precondition for the rollout of projects. Beyond that, John Concato, Nirav Shah, and Horwitz (2000) argue that, in practice, RCTs do not provide useful information beyond what can be learned from well-designed and carefully interpreted observational studies.

5. *Where Should We Go from Here?*

Cartwright (2007b) maintains a useful distinction between "hunting" causes and "using" them, and this section is about the use of randomized controlled trials for policy. Here I address the issue of generalizability or external validity—as opposed to internal validity as discussed in the previous section—grounds on which development RCTs are sometimes criticized (see, for example, Dani Rodrik 2009). We need to know when we can use local results, from instrumental variables, from RCTs, or from nonexperimental analyses, in contexts other than those in which they were obtained.

There are certainly cases in both medicine and economics where an RCT has had a major effect on the way that people think beyond the original local context of the trial. In the recent development literature, my favorite is Raghabendra Chattopadhyay and Duflo's (2004) study of women leaders in India. The Government of India randomly selected some *panchayats* and forced them to have female leaders, and the paper explores the differences in outcomes between such villages and others with male leaders. There is a theory (of sorts) underlying these experiments—the development community had briefly adopted the view that a key issue in development was the empowerment of women (or perhaps just giving them "voice") and, if this was done, more children would be educated, more money would be spent on food and on health, and so on. Women are altruistic and men are selfish. Chattopadhyay and Duflo's analysis of the Indian government's experiments shows that this is most likely wrong; I say most likely because, as with all experiments, the mechanisms are unclear. It is possible, for example, that women do indeed want the socially desirable outcomes but are unable to obtain them in a male dominated society, even when women are nominally in power. Even so, this study reversed previously held general beliefs. There are also many examples in medicine where knowledge of the mean treatment effect among the treated from a trial, even with some allowance for practical problems, has reversed previously held beliefs (see Davey Smith and Shah Ebrahim 2002, who note that "Observational studies propose, RCTs dispose").

Yet I also believe that RCTs of "what works," even when done without error or contamination, are unlikely to be helpful for policy, or to move beyond the local, unless they tell us something about why the program worked, something to which they are often neither targeted nor well-suited. Some of the issues are familiar and are widely discussed in the literature. Actual policy is always likely to be different from the experiment, for example because there are general equilibrium effects that operate on a large scale that are absent in a pilot, or because the outcomes are different when *everyone* is covered by the treatment rather than just a selected group of experimental subjects who are not representative of the population to be covered by the policy. Small development projects that help a few villagers or a few villages may not attract the attention of corrupt public officials because it is not worth their while to undermine or exploit them, yet they would do so as soon as any attempt were made to scale up. The scientists who run the experiments are likely to do so more carefully and conscientiously than would the bureaucrats in charge of a full scale operation. In consequence, there is no guarantee that the policy tested by the RCT will have the same effects as in the trial, even on the subjects included in the trial or in the population from which the trialists were selected. For an RCT to produce "useful knowledge" beyond its local context, it must illustrate some general tendency, some effect that is the result of mechanism that is likely to apply more broadly.

It is sometimes argued that skepticism about generalizability is simply "a version of David Hume's famous demonstration of the lack of a rational basis for induction" (Banerjee 2005, p. 4341). But what is going on here is often a good deal more mundane. Worrall (2007, p. 995) responds to the same argument with the following: "One example is the drug benoxaprophen (trade name: Opren), a nonsteroidal inflammatory treatment for arthritis and musculo-skeletal pain. This passed RCTs (explicitly restricted to 18 to 65 year olds) with flying colours. It is however a fact that musculo-skeletal pain predominately afflicts the elderly. It turned out, when the (on average older) 'target population' were given Opren, there were a significant number of deaths from hepato-renal failure and the drug was withdrawn."

In the same way, an educational protocol that was successful when randomized across villages in India holds many things constant that would not be constant if the program were transported to Guatemala or Vietnam, even as a randomized controlled trial, let alone when enacted into policy (Cartwright 2010). These examples demonstrate a failure to control for relevant factors or nonrandom participation in the trial, not the general impossibility of induction. RCTs, like nonexperimental results, cannot automatically be extrapolated outside the context in which they were obtained.

Perhaps the most famous randomization in development economics is Progresa (now *Oportunidades*) in Mexico, a conditional cash transfer scheme in which welfare benefits to parents were paid conditional on their children attending schools and clinics. Angrist and Pischke (2010) approvingly quote Paul Gertler's statement that "Progresa is why now thirty countries worldwide have conditional cash transfer programs," and there is no doubt that the spread of Progresa depended on the fact that its successes were supported by a randomized evaluation. Yet it is unclear that this wholesale imitation is a good thing. Santiago Levy (2006), the architect of Progresa, notes that the Mexican scheme cannot simply be exported to other countries if, for example, those countries have a preexisting antipoverty program with which conditional cash transfers might not fit, or if they do not have the capacity to meet the additional demand for education or healthcare, or if the political support is absent. Incentivizing parents to take their children to clinics will not improve child health if there are no clinics to serve them, a detail that can easily be overlooked in the enthusiasm for the credibility of the Mexican evaluation.

Pawson and Tilley (1997) argue that it is the combination of mechanism and context that generates outcomes and that, without understanding that combination, scientific progress is unlikely. Nor can we safely go from experiments to policy. In economics, the language would refer to theory, building models, and tailoring them to local conditions. Policy requires a causal model; without it, we cannot understand the welfare consequences of a policy, even a policy where causality is established and that is proven to work on its own terms. Banerjee (2007a) describes an RCT by Duflo, Rema Hanna, and Stephen Ryan (2008) as "a new economics being born." This experiment used cameras to monitor and prevent teacher absenteeism in villages in the Indian state of Rajasthan. Curiously, Pawson and Tilley (1997, pp. 78–82) use the example of cameras (to deter crime in car parks) as one of their running examples. They note that cameras do not, in and of themselves, prevent crime because they do not make it impossible to break into a car. Instead, they depend on triggering a series of behavioral changes. Some of those changes show positive experimental outcomes—crime is down in the car parks with cameras—but are undesirable, for example because crime is shifted to other car parks or because the cameras change the mix of patrons of the car park. There are also cases where the experiment fails but has beneficial effects. It would not be difficult to construct similar arguments for the cameras in the Indian schools, and welfare conclusions cannot be supported unless we understand the behavior of teachers, pupils, and their parents. Duflo, Hanna, and Ryan (2008) understand this and use their experimental results to construct a model of teacher behavior. Other papers that use structural models to interpret experimental results include Petra E. Todd and Kenneth I. Wolpin (2006) and Orazio Attanasio, Costas Meghir, and Ana Santiago (2005); these and the other studies reviewed in Todd and Wolpin (forthcoming) are surely a good avenue for future explanation.

Cartwright (2007a) draws a contrast between the rigor applied to establish internal validity—to establish the gold standard status of RCTs—and the much looser arguments that are used to defend the transplantation of the experimental results to policy. For example, running RCTs to find out whether a project works is often defended on the grounds that the experimental project is like the policy that it might support. But the "like" is typically argued by an appeal to similar circumstances, or a similar environment, arguments that depend entirely on observable variables. Yet controlling for observables is the key to the matching estimators that are one of the main competitors for RCTs and that are typically rejected by the advocates of RCTs on the grounds that RCTs control not only for the things that we observe but things that we cannot. As Cartwright notes, the validity of evidence-based policy depends on the *weakest* link in the chain of argument and evidence, so that by the time we seek to use the experimental results, the advantage of RCTs over matching or other econometric methods has evaporated. In the end, there is no substitute for careful evaluation of the chain of evidence and reasoning by people who have the experience and expertise in the field. The demand that experiments be theory-driven is, of course, no guarantee of success, though the lack of it is close to a guarantee of failure.

It is certainly not always obvious how to combine theory with experiments. Indeed, much of the interest in RCTs—and in instrumental variables and other econometric techniques that mimic random allocation—comes from a deep skepticism of economic theory, and impatience with its ability to deliver structures that seem at all helpful in interpreting reality. Applied and theoretical economists seem to be further apart now than at any period in the last quarter century. Yet failure to reintegrate is hardly an option because without it there is no chance of long-term scientific progress or of maintaining and extending the results of experimentation. RCTs that are not theoretically guided are unlikely to have more than local validity, a warning that applies equally to nonexperimental work. In Angus Deaton (forthcoming), where I develop these arguments further, I discuss a number of examples of nonexperimental work in economic development where theories are developed to the point where they are capable of being tested on nonexperimental data, with the results used to refute, refine, or further develop the theory. Randomized experiments, which allow the researcher to induce controlled variance, should be a powerful tool in such programs, and make it possible to construct tests of theory that might otherwise be difficult or impossible. The difference is not in the methods, experimental and nonexperimental, but in what is being investigated, projects on the one hand, and mechanisms on the other.

One area in which this is already happening is in behavioral economics, and the merging of economics and psychology, whose own experimental tradition is clearly focused on behavioral regularities. The experiments reviewed in Steven D. Levitt and John A. List (2008), often involving both economists and psychologists, cover such issues as loss aversion, procrastination, hyperbolic discounting, or the availability heuristic—all of which are examples of behavioral mechanisms that promise applicability beyond the specific experiments. There also appears to be a good deal of convergence between this line of work, inspired by earlier experimental traditions in economic theory and in psychology, and the most recent work in development. Instead of using experiments to evaluate projects, looking for which projects work, this development work designs its experiments to test predictions of theories that are generalizable to other situations. Without any attempt to be comprehensive, some examples are Dean Karlan and Jonathan Zinman (2008), who are concerned with the price elasticity of the demand for credit; Marianne Bertrand et al. (2010), who take predictions about the importance of context from the psychology laboratory to the study of advertising for small loans in South Africa; Duflo, Kremer, and Jonathan Robinson (2009), who construct and test a behavioral model of procrastination for the use of fertilizers by small farmers in Kenya; and Xavier Giné, Karlan, and Zinman (2009), who use an experiment in the Philippines to test the efficacy of a smoking-cessation product designed around behavioral theory. In all of this work, the project, when it exists at all, is an embodiment of the theory that is being tested and refined, not the object of evaluation in its own right, and the field experiments are a bridge between the laboratory and the analysis of "natural" data (List 2006). The collection of purpose-designed data and the use of randomization often make it easier to design an acid test that can be more difficult to construct without them. If we are lucky, this work will provide the sort of behavioral realism that has been lacking in much of economics while, at the same time, identifying and allowing us to retain the substantial parts of existing economic theory that are genuinely useful.

In this context, it is worth looking back to the previous phase of experimentation in economics that started with the New Jersey income tax experiments. A rationale for these experiments is laid out in Guy H. Orcutt and Alice G. Orcutt (1968) in which the vision is a formal model of labor supply with the experiments used to estimate its parameters. By the early 1990s, however, experimentation had moved on to a "what works" basis, and Manski and Irwin Garfinkel (1992), surveying the experience, write "there is, at present, no basis for the popular belief that extrapolation from social experiments is less problematic than extrapolation from observational data. As we see it, the recent embrace of reduced-form social experimentation to the exclusion of structural evaluation based on observational data is not warranted." Their statement still holds good, and it would be worth our while trying to return to something like Orcutt and Orcutt's vision in experimental work, as well as to reestablishing the relevance and importance of theory-driven nonexperimental work. The more recent Moving to Opportunity experiment is another example of an experiment that was more successful as a black-box test of "what works," in this case giving housing vouchers to a small segment of the most disadvantaged population, than it was in illuminating general long-standing questions about the importance of neighborhoods (Sampson 2008). Sampson concludes his review of the Moving to Opportunity experiment with "a plea for the old-fashioned but time-proven benefits of theoretically motivated descriptive research and synthetic analytical efforts" (p. 227).

Finally, I want to return to the issue of "heterogeneity," a running theme in this paper. Heterogeneity of responses first appeared in section 2 as a technical problem for instrumental variable estimation, dealt with in the literature by local average treatment estimators. Randomized controlled trials provide a method for estimating quantities of interest in the presence of heterogeneity, and can therefore be seen as another technical solution for the "heterogeneity problem." They allow estimation of mean responses under extraordinarily weak conditions. But as soon as we deviate from ideal conditions and try to correct the randomization for inevitable practical difficulties, heterogeneity again rears its head, biasing estimates, and making it difficult to interpret what we get. In the end, the technical fixes fail and compromise our attempts to learn from the data. What this should tell us is that the heterogeneity is *not* a technical problem but a symptom of something deeper, which is the failure to specify causal models of the processes we are examining. Technique is never a substitute for the business of doing economics.

Perhaps the most successful example of learning about economic development is the Industrial Revolution in Britain, recently described by Joel Mokyr (2009). Much of the learning was indeed accomplished by trial and error—RCTs were yet to be invented—though the practical and often uneducated people who did the experimentation were constantly in contact with leading scientists, often through hundreds of local societies dedicated to the production of "useful knowledge" and its application to the benefit of mankind (see also Roy Porter 2000). Mokyr sees the Enlightenment as a necessary precondition for the revolution and of the escape from poverty and disease. Yet the application and development of scientific ideas was uneven, and progress was faster where there was less heterogeneity, for example in chemical or mechanical processes, which tended to work in much the same way everywhere, than where heterogeneity was important, as in agriculture, where soils and farms differed by the mile (Mokyr 2009, p. 192). Even so, as with sanitation, progress was often made without a correct understanding of mechanisms and with limited or no experimentation; Richard J. Murnane and Richard R. Nelson (2007) argue that the same is frequently true in modern medicine. In the end, many problems were simply too hard to be solved without theoretical guidance, which in areas such as soil chemistry, or the germ theory of disease, lay many decades in the future. It took scientific understanding to overcome the heterogeneity of experience which ultimately defeats trial and error. As was the case then, so now, and I believe that we are unlikely to banish poverty in the modern world by trials alone, unless those trials are guided by and contribute to theoretical understanding.

References

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review*, 91(5): 1369–1401.

Altman, Douglas G. 1998. "Within Trial Variation—A False Trail?" *Journal of Clinical Epidemiology*, 51(4): 301–03.

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review*, 80(3): 313–36.

Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics*, 114(2): 533–75.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con Out of Econometrics." *Journal of Economic Perspectives*, 24(2): 3–30.

Attanasio, Orazio, Costas Meghir, and Ana Santiago. 2005. "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate Progresa." Unpublished.

Banerjee, Abhijit Vinayak. 2005. "'New Development Economics' and the Challenge to Theory." *Economic and Political Weekly*, 40(40): 4340–44.

Banerjee, Abhijit Vinayak. 2007a. "Inside the Machine: Toward a New Development Economics." *Boston Review*, 32(2): 12–18.

Banerjee, Abhijit Vinayak. 2007b. *Making Aid Work*. Cambridge and London: MIT Press.

Banerjee, Abhijit Vinayak, and Ruimin He. 2008. "Making Aid Work." In *Reinventing Foreign Aid*, ed. William Easterly, 47–92. Cambridge and London: MIT Press.

Barro, Robert J. 1998. *Determinants of Economic Growth: A Cross-Country Empirical Study*. Cambridge and London: MIT Press.

Barro, Robert J., and Xavier Sala-i-Martin. 1995. *Economic Growth*. New York; London and Montreal: McGraw-Hill.

Bauer, P. T. 1971. *Dissent on Development: Studies and Debates in Development Economics*. London: Weidenfeld and Nicolson.

Bauer, P. T. 1981. *Equality, the Third World and Economic Delusion*. Cambridge and London: Harvard University Press.

Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman. 2010. "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment." *Quarterly Journal of Economics*, 125(1): 263–305.

Blundell, Richard, and Monica Costa Dias. 2009. "Alternative Approaches to Evaluation in Empirical Microeconomics." *Journal of Human Resources*, 44(3): 565–640.

Boone, Peter. 1996. "Politics and the Effectiveness of Foreign Aid." *European Economic Review*, 40(2): 289–329.

Burnside, Craig, and David Dollar. 2000. "Aid, Policies, and Growth." *American Economic Review*, 90(4): 847–68.

Card, David. 1999. "The Causal Effect of Education on Earnings." In *Handbook of Labor Economics, Volume 3A*, ed. Orley Ashenfelter and David Card, 1801–63. Amsterdam; New York and Oxford: Elsevier Science, North-Holland.

Cartwright, Nancy. 2007a. "Are RCTs the Gold Standard?" *Biosocieties*, 2(1): 11–20.

Cartwright, Nancy. 2007b. *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge and New York: Cambridge University Press.

Cartwright, Nancy. 2010. "What Are Randomised Controlled Trials Good For?" *Philosophical Studies*, 147(1): 59–70.

Chattopadhyay, Raghabendra, and Esther Duflo. 2004. "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India." *Econometrica*, 72(5): 1409–43.

Clemens, Michael, Steven Radelet, and Rikhil Bhavnani. 2004. "Counting Chickens When They Hatch: The Short Term Effect of Aid on Growth." Center for Global Development Working Paper 44.

Concato, John, Nirav Shah, and Ralph I. Horwitz. 2000. "Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs." *New England Journal of Medicine*, 342(25): 1887–92.

Cox, David R. 1958. *Planning of Experiments*. New York: Wiley.

Dalgaard, Carl-Johan, and Henrik Hansen. 2001. "On Aid, Growth and Good Policies." *Journal of Development Studies*, 37(6): 17–41.

Dalgaard, Carl-Johan, Henrik Hansen, and Finn Tarp. 2004. "On the Empirics of Foreign Aid and Growth." *Economic Journal*, 114(496): F191–216.

Davey Smith, George, and Shah Ebrahim. 2002. "Data Dredging, Bias, or Confounding: They Can All Get You into the *BMJ* and the Friday Papers." *British Medical Journal*, 325: 1437–38.

Davey Smith, George, and Matthias Egger. 1998. "Incommunicable Knowledge? Interpreting and Applying the Results of Clinical Trials and Meta-analyses." *Journal of Clinical Epidemiology*, 51(4): 289–95.

Deaton, Angus. Forthcoming. "Understanding the Mechanisms of Economic Development." *Journal of Economic Perspectives*.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2008. "Returns to Capital in Microenterprises: Evidence from a Field Experiment." *Quarterly Journal of Economics*, 123(4): 1329–72.

Duflo, Esther. 2004. "Scaling Up and Evaluation." In *Annual World Bank Conference on Development Economics, 2004: Accelerating Development*, ed. François Bourguignon and Boris Pleskovic, 341–69. Washington, D.C.: World Bank; Oxford and New York: Oxford University Press.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2008. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics, Volume 4*, ed. T. Paul Schultz and John Strauss, 3895–3962. Amsterdam and Oxford: Elsevier, North-Holland.

Duflo, Esther, Rema Hanna, and Stephen Ryan. 2008. "Monitoring Works: Getting Teachers to Come to School." Center for Economic and Policy Research Discussion Paper 6682.

Duflo, Esther, Michael Kremer, and Jonathan Robinson. 2009. "Nudging Farmers to Use Fertilizer: Evidence from Kenya." National Bureau of Economic Research Working Paper 15131.

Duflo, Esther, and Rohini Pande. 2007. "Dams." *Quarterly Journal of Economics*, 122(2): 601–46.

Easterly, William. 2006. *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*. New York: Penguin Press.

Easterly, William, ed. 2008. *Reinventing Foreign Aid*. Cambridge and London: MIT Press.

Easterly, William. 2009. "Can the West Save Africa?" *Journal of Economic Literature*, 47(2): 373–447.

Easterly, William, Ross Levine, and David Roodman. 2004. "Aid, Policies, and Growth: Comment." *American Economic Review*, 94(3): 774–80.

Fisher, Ronald A. 1935. *The Design of Experiments*, Eighth edition. New York: Hafner, 1960.

Freedman, David A. 2005. *Statistical Models: Theory and Practice*. Cambridge and New York: Cambridge University Press.

Freedman, David A. 2006. "Statistical Models for Causation: What Inferential Leverage Do They Provide?" *Evaluation Review*, 30(6): 691–713.

Freedman, David A. 2008. "On Regression Adjustments to Experimental Data." *Advances in Applied Mathematics*, 40(2): 180–93.

Giné, Xavier, Dean Karlan, and Jonathan Zinman. 2009. "Put Your Money Where Your Butt Is: A Commitment Contract for Smoking Cessation." World Bank Policy Research Working Paper 4985.

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz. 2004. "Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya." *Journal of Development Economics*, 74(1): 251–68.

Groopman, Jerome. 2009. "Diagnosis: What Doctors Are Missing." *New York Review of Books*, 56(17), November 5th.

Guillaumont, Patrick, and Lisa Chauvet. 2001. "Aid and Performance: A Reassessment." *Journal of Development Studies*, 37(6): 66–92.

Hansen, Henrik, and Finn Tarp. 2000. "Aid Effectiveness Disputed." *Journal of International Development*, 12(3): 375–98.

Hansen, Henrik, and Finn Tarp. 2001. "Aid and Growth Regressions." *Journal of Development Economics*, 64(2): 547–70.

Heckman, James J. 1992. "Randomization and Social Policy Evaluation." In *Evaluating Welfare and Training Programs*, ed. Charles F. Manski and Irwin Garfinkel, 201–30. Cambridge and London: Harvard University Press. Available as National Bureau of

Economic Research Technical Working Paper 107.

Heckman, James J. 1997. "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations." *Journal of Human Resources*, 32(3): 441–62.

Heckman, James J. 2000. "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective." *Quarterly Journal of Economics*, 115(1): 45–97.

Heckman, James J., and Jeffrey A. Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives*, 9(2): 85–110.

Heckman, James J., and Sergio Urzua. 2009. "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify." National Bureau of Economic Research Working Paper 14706.

Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *Review of Economics and Statistics*, 88(3): 389–432.

Heckman, James J., and Edward Vytlacil. 1999. "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects." *Proceedings of the National Academy of Sciences*, 96(8): 4730–34.

Heckman, James J., and Edward Vytlacil. 2007. "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments." In *Handbook of Econometrics, Volume 6B*, ed. James J. Heckman and Edward E. Leamer, 4875–5143. Amsterdam and Oxford: Elsevier, North-Holland.

Horwitz, Ralph I., Burton H. Singer, Robert W. Makuch, and Catherine M. Viscoli. 1996. "Can Treatment That is Helpful on Average Be Harmful to Some Patients? A Study of the Conflicting Information Needs of Clinical Inquiry and Drug Regulation." *Journal of Clinical Epidemiology*, 49(4): 395–400.

Horwitz, Ralph I., Burton H. Singer, Robert W. Makuch, and Catherine M. Viscoli. 1997. "On Reaching the Tunnel at the End of the Light." *Journal of Clinical Epidemiology*, 50(7): 753–55.

Hoxby, Caroline M. 2000. "Does Competition among Public Schools Benefit Students and Taxpayers?" *American Economic Review*, 90(5): 1209–38.

Imbens, Guido W. 2009. "Better Late than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." National Bureau of Economic Research Working Paper 14896.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–75.

International Initiative for Impact Evaluation (3IE). 2008. http://www.3ieimpact.org/.

Jurajda, Štěpán, and Daniel Münich. 2009. "Admission to Selective Schools, Alphabetically." http://home.cerge-ei.cz/jurajda/alphabet07.pdf.

Kanbur, Ravi. 2001. "Economic Policy, Distribution and Poverty: The Nature of Disagreements." *World Development*, 29(6): 1083–94.

Karlan, Dean, and Jonathan Zinman. 2008. "Credit Elasticities in Less-Developed Economies: Implications for Microfinance." *American Economic Review*, 98(3): 1040–68.

Leamer, Edward E. 1985. "Vector Autoregressions for Causal Inference?" *Carnegie–Rochester Conference Series on Public Policy*, 22: 255–303.

Lee, David S., and Thomas Lemieux. 2009. "Regression Discontinuity Designs in Economics." National Bureau of Economic Research Working Paper 14723.

Lensink, Robert, and Howard White. 2001. "Are There Negative Returns to Aid?" *Journal of Development Studies*, 37(6): 42–65.

Leonhardt, David. 2008. "Making Economics Relevant Again." *New York Times*, February 20.

Levitt, Steven D., and John A. List. 2008. "Field Experiments in Economics: The Past, the Present, and the Future." National Bureau of Economic Research Working Paper 14356.

Levy, Santiago. 2006. *Progress against Poverty: Sustaining Mexico's Progresa-Oportunidades Program.* Washington, D.C.: Brookings Institution Press.

List, John A. 2006. "Field Experiments: A Bridge between Lab and Naturally Occurring Data." *B.E. Journal of Economic Analysis and Policy: Advances in Economic Analysis and Policy*, 6(2).

Mankiw, N. Gregory, David Romer, and David N. Weil. 1992. "A Contribution to the Empirics of Economic Growth." *Quarterly Journal of Economics*, 107(2): 407–37.

Manski, Charles F. 1996. "Learning About Treatment Effects from Experiments with Random Assignment of Treatments." *Journal of Human Resources*, 31(4): 709–33.

Manski, Charles F., and Irwin Garfinkel. 1992. "Evaluating Welfare and Training Programs: Introduction." In *Evaluating Welfare and Training Programs*, ed. Charles F. Manski and Irwin Garfinkel, 1–22. Cambridge and London: Harvard University Press.

McCleary, Rachel M., and Robert J. Barro. 2006. "Religion and Economy." *Journal of Economic Perspectives*, 20(2): 49–72.

McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics*, 142(2): 698–714.

Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica*, 72(1): 159–217.

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. "Economic Shocks and Civil Conflict: An Instrumental Variables Approach." *Journal of Political Economy*, 112(4): 725–53.

Mokyr, Joel. 2009. *The Enlightened Economy: An Economic History of Britain 1700–1850*. New Haven and London: Yale University Press.

Murnane, Richard J., and Richard R. Nelson. 2007. "Improving the Performance of the Education Sector: The Valuable, Challenging, and Limited Role of Random Assignment Evaluations." *Economics of Innovation and New Technology*, 16(5): 307–22.

Orcutt, Guy H., and Alice G. Orcutt. 1968. "Incentive and Disincentive Experimentation for Income Maintenance Policy Purposes." *American Economic Review*, 58(4): 754–72.

Pawson, Ray, and Nick Tilley. 1997. *Realistic Evaluation*. London and Thousand Oaks, Calif.: Sage Publications.

Peto, Richard, Rory Collins, and Richard Gray. 1995. "Large-Scale Randomized Evidence: Large, Simple Trials and Overviews of Trials." *Journal of Clinical Epidemiology*, 48(1): 23–40.

Pogge, Thomas. 2005. "World Poverty and Human Rights." *Ethics and International Affairs*, 19(1): 1–7.

Porter, Roy. 2000. *The Creation of the Modern World: The Untold Story of the British Enlightenment*. New York and London: Norton.

Poverty Action Lab. 2007. "Clinton Honors Global Deworming Effort." http://www.povertyactionlab.org/deworm/.

Rajan, Raghuram G., and Arvind Subramanian. 2008. "Aid and Growth: What Does the Cross-Country Evidence Really Show?" *Review of Economics and Statistics*, 90(4): 643–65.

Reiss, Peter C., and Frank A. Wolak. 2007. "Structural Econometric Modeling: Rationales and Examples from Industrial Organization." *Handbook of Econometrics, Volume 6A*, ed. James J. Heckman and Edward E. Leamer, 4277–4415. Amsterdam and Oxford: Elsevier, North-Holland.

Rodrik, Dani. 2009. "The New Development Economics: We Shall Experiment, but How Shall We Learn?" In *What Works in Development? Thinking Big and Thinking Small*, ed. Jessica Cohen and William Easterly, 24–47. Washington, D.C.: Brookings Institution Press.

Roodman, David. 2007. "The Anarchy of Numbers: Aid, Development, and Cross-Country Empirics." *World Bank Economic Review*, 21(2): 255–77.

Sachs, Jeffrey D. 2005. *The End of Poverty: Economic Possibilities for Our Time*. New York: Penguin.

Sachs, Jeffrey D. 2008. *Common Wealth: Economics for a Crowded Planet*. New York: Penguin Press.

Sampson, Robert J. 2008. "Moving to Inequality: Neighborhood Effects and Experiments Meet Social Structure." *American Journal of Sociology*, 114(1): 189–231.

Sanbonmatsu, Lisa, Jeffrey R. Kling, Greg J. Duncan, and Jeanne Brooks-Gunn. 2006. "Neighborhoods and Academic Achievement: Results from the Moving to Opportunity Experiment." *Journal of Human Resources*, 41(4): 649–91.

Senn, Stephen, and Frank Harrell. 1997. "On Wisdom after the Event." *Journal of Clinical Epidemiology*, 50(7): 749–51.

Sims, Christopher A. 2010. "But Economics Is Not an Experimental Science." *Journal of Economic Perspectives*, 24(2): 59–68.

Singer, Peter. 2004. *One World: The Ethics of Globalization*. New Haven and London: Yale University Press, 2002.

The Lancet. 2004. "The World Bank Is Finally Embracing Science: Editorial." *The Lancet*, 364: 731–32.

Todd, Petra E., and Kenneth I. Wolpin. 2006. "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility." *American Economic Review*, 96(5): 1384–1417.

Todd, Petra E., and Kenneth I. Wolpin. Forthcoming. "Structural Estimation and Policy Evaluation in Developing Countries." *Annual Review of Economics*.

Urquiola, Miguel, and Eric Verhoogen. 2009. "Class-Size Caps, Sorting, and the Regression-Discontinuity Design." *American Economic Review*, 99(1): 179–215.

van den Berg, Gerard. 2008. "An Economic Analysis of Exclusion Restrictions for Instrumental Variable Estimation." Unpublished.

Wennberg, David E., F. L. Lucas, John D. Birkmeyer, Carl E. Bredenberg, and Elliott S. Fisher. 1998. "Variation in Carotid Endarterectomy Mortality in the Medicare Population: Trial Hospitals, Volume, and Patient Characteristics." *Journal of the American Medical Association*, 279(16): 1278–81.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge and London: MIT Press.

World Bank. 2008. "Spanish Impact Evaluation Fund." http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,contentMDK:21419502~menuPK:384336~pagePK:148956~piPK:216618~theSitePK:384329,00.html.

Worrall, John. 2007. "Evidence in Medicine and Evidence-Based Medicine." *Philosophy Compass*, 2(6): 981–1022.