SCHOOL OF ECONOMICS AND MANAGEMENT
FACULTY OF SOCIAL SCIENCES
AARHUS UNIVERSITY

CREATES
Center for Research in Econometric Analysis of Time Series

CREATES Research Paper 2011-25

Field Experiments in Economics: Comment on an article by Levitt and List

Stephen T. Ziliak

School of Economics and Management
Aarhus University
Bartholins Allé 10, Building 1322, DK-8000 Aarhus C
Denmark

Field Experiments in Economics:
Comment on an article by Levitt and List

by Stephen T. Ziliak

Trustee and Professor of Economics
Roosevelt University
430 S. Michigan Ave
Chicago, IL 60605

http://sites.roosevelt.edu/sziliak
http://stephentziliak.com
email: sziliak@roosevelt.edu

June 2011

Abstract: In an article titled "Field Experiments in Economics: The Past, the Present, and the Future," Levitt and List (2009) make three important claims about the history, philosophy, and future of field experiments in economics. They claim that field experiments in economics began in the 1920s and 1930s, in agricultural work by Neyman and Fisher. Second, they claim that artificial randomization is the *sine qua non* of good experimental design; they claim that randomization is the only valid justification for use of Student's test of significance. Finally, they claim that the theory of the firm will be advanced by economists doing randomized controlled trials (RCTs) for private sector firms. Several areas of economics, for example the development economics of Banerjee and Duflo, have been influenced by the article, despite the absence of historical and methodological review. This comment seeks to fill that gap in the literature. Student has, it is found, priority over Fisher and Neyman; he compared balanced and random designs in the field—on crops from barley to timber—from 1905 to 1937. The power and efficiency of balanced over random designs - discovered by Student and confirmed by Pearson, Neyman, Jeffreys, and others adopting a decision-theoretic and/or Bayesian approach - is not mentioned by Levitt and List. Neglect of Student is especially regrettable, for he showed in his job as Head Brewer of Guinness that artificial randomization is neither necessary nor sufficient for improving efficiency, identifying causal relationships, or discovering economically significant differences. One way forward is to take a step backwards, from Fisher to Student.

JEL classification: B1, C9, C93 Keywords: field experiments, balanced, random

Field Experiments in Economics:
Comment on an article by Levitt and List

By Stephen T. Ziliak[1]

A model whose faults you can all too easily acquire is sure to mislead you.
Horace (20 B.C.)

Randomization is a metaphor and not an ideal or "gold standard".
Heckman and Vytlacil (2007)

1. Introduction

In an article titled "Field Experiments in Economics: The Past, the Present, and the Future," Levitt and List (2009) make three important claims about the history, philosophy, and future of field experiments in economics.

Claim number one says that field experiments in economics began "in the 1920s and 1930s" (p. 1), in agricultural research by the mathematical statisticians' Jerzy Neyman and Ronald A. Fisher (Levitt and List 2009, pp. 1, 3-5).

Claim number two says that artificial randomization (Levitt and List 2009, p. 3) is the *sine qua non* of good experimental design in economics. The claim here is that randomization of design— the random assignment of treatments, controls, and/or varieties to experimental units— is the only "valid" (Levitt and List, p. 3) justification for using Student's test of significance (Levitt and List, p. 3).

An example of artificial randomization is the use of dice, shuffled cards, or some other random number generator to allocate treatments and controls to an experimental unit. A potato farmer, for example, may wish to test the hypothesis that, other things equal, crop yield is higher when crops are fertilized—the unfertilized crops serving as controls. For proving the validity of

[1] I thank audiences at the annual meetings of the American Association for Cancer Research (Washington, DC, 2010) and of the American Statistical Association (Washington, DC, 2009 and Chicago chapter, 2009) and at CREATES (Aarhus University), the John Marshall Law School, the National Institutes of Health (Division of Biomedical Computing), the University of Illinois-Chicago Department of Epidemiology and Biostatistics, the University of Illinois-Urbana-Champaign Department of Agricultural and Consumer Economics, and at the University of Kentucky: the departments of Economics, Statistics, Psychology, and Political Science. For helpful comments and discussion of the basic issues I thank the editor, two anonymous referees, and: Doug Altman, Jim Conybear, Jim DeLeo, Hakan Demirtas, Tom Engsted, Sally Freels, Steve Goodman, M.D., Ph.D., Craig Gundersen, Tim Harford, Jack Lee, Carl Leonard, Pedro Lowenstein, M.D., Ph.D., Deirdre McCloskey, Jacques Kibambe Ngoie, Ozgur Orhangazi, Dani Rodrik, Allan Wurtz, the late Arnold Zellner (1927-2010), and Jim Ziliak. For valuable assistance in the archives and permission to quote, I kindly thank principals of the Guinness Archives (Diageo) and of the Special Collections Library, University College London. Any errors are my own.

randomization in the allocation of treatments and controls the authors credit "Splawa-Neyman (1923 [1990])" and "Fisher and McKenzie (1923)" [sic] (Levitt and List, pp. 2-4)—these articles, Levitt and List claim, established the "experimental foundation" (Levitt and List, p. 1) which they equate with randomization.[2]

Their third and final claim about the past, present, and future of field experiments in economics begins with an assertion that the "current generation" (Levitt and List, p. 7) has seen "much deeper" (Levitt and List, p. 15) and "broader" (p. 15) than "previous generations" (p. 15). Since the late 1990s, they argue, a new crop of published articles exceed the century of theory and results pioneered by William S. Gosset (1876-1937) aka Student, and by Fisher, Neyman, Pearson, Jeffreys, Yates, and others down to Zellner and Heckman. Thus in their third major claim, both empirical economics and the theory of the firm (pp. 2-15) will advance as, they assume, medicine and pharmacology have advanced, as more and more economists conduct randomized trials with private sector firms, a "win-win" (p. 15) opportunity, they claim. [3]

"We view such partnerships as permitting the analyst a unique inside view that will not only provide a glimpse into the decision-making black box, but permit a deeper empirical exploration into problems that excite economists, practitioners, and policymakers" (L and L, p. 2).

The success of "Field Experiments in Economics: The Past, the Present, and the Future" depends on the degree to which each of its three main claims is—or might be—theoretically and empirically established.

To date no study has endeavored to assess their history and philosophy of field experiments in economics. This is unfortunate but, on second thought, not surprising. As Deaton (2007, p. 25) observes, "[t]he movement in favor of RCTs [randomized controlled trials] is currently very successful" in development economics and "increasingly", Levitt and List note, "in the economics literature" at large (L and L, pp. 2, 15-18). "I am a huge fan of randomized trials," Varian (2011) wrote in *The Economist*, "they are well worth doing since they are the gold standard for causal inference".

Deaton's (2007) appraisal of RCTs is mostly negative. Economists at the frontier of the "burgeoning literature" (L and L, p. 2)—for instance, Duflo, Banerjee, and others at the Abdul Latif Jameel Poverty Action Lab (J-PAL (2010))—are, Deaton finds, simply assuming the validity of randomization, not proving it.[4]

Recently Duflo and other economists have turned to Levitt and List (2009) to justify randomization and its place in the history and philosophy of economic science (J-PAL, 2010). "The thoroughness of Fisher's insights are exemplified by this passage", Levitt and List write, "concerning what constituted a valid randomization scheme for completely randomized blocks"

---

[2] Splawa-Neyman and Neyman are the same person—the mathematical statistician, Jerzy Neyman (Reid 1982). Neyman shortened his name in the late 1920s when he moved from Poland to England to accept a temporary position at University College London.

[3] But see the revised CONSORT statement on randomized controlled trials in medicine by Altman, Schultz, Moher, Egger, Davidoff, Elbourne, Gotzsche, and Lang (2001) and also Rothman, Greenland, and Lang (2008); for discussion of Student's pioneering comparisons of random and balanced experiments see Ziliak (2011a, 2011b, 2010, 2008) and Ziliak and McCloskey (2008, chps. 1, 20-23).

[4] Contrast Rodrik (2008); Bruhn and McKenzie (2009).

(Levitt and List 2009, p. 3). Here is the passage that is supposed to reveal the thoroughness of Fisher's (1935, p. 26) insights:

> The validity of our estimate of error for this purpose is guaranteed by the provision that any two plots, not in the same block, shall have the same probability of being treated alike, and the same probability of being treated differently in each of the ways in which this is possible (quoted by Levitt and List, p. 3).

Levitt and List - following Cochrane (1976), Rubin (1990), and Street (1990), on whom their views are largely dependent - assume rather than prove that Fisher "laid the experimental foundation" (L and L, abstract) with "completely randomized blocks" (p. 3). The claims sound persuasive, and highly scientific. And they make numerous appeals to experimental history and methods descending from the Fisher School. Unfortunately, most economists do not know how to judge these claims. Most economists are ill-equipped to judge their sweeping methodological and historical assertions (see also: Harrison and List (2004)).

This comment draws on a century of theory and evidence not examined by Levitt and List—bringing into focus each of their three main claims: for example, Gosset (1904, 1905b: in Pearson (1939)), Student (1908, 1911, 1923, 1936, 1938, 1942), Wood and Stratton (1910), Mercer and Hall (1911), Pearson (1938, 1990), Neyman, Iwaszkiewicz, and Kolodziejczyk (1935), Neyman (1938), Neyman and Pearson (1938), Savage (1954, 1976), Jeffreys (1939 [1961]), Kruskal (1980), Zellner and Rossi (1986), Heckman and Vytlacil (2007), Press (2003), Carson, Louviere, and Wasi (2009), Bruhn and McKenzie (2009), and others.[5]

Neglect of Student (1911, 1923, 1938) is especially regrettable, for Student showed by repeating experiments in a profit-seeking environment – in his job as Head Experimental Brewer and finally as Head Brewer of Guinness, Dublin and Park Royal - that artificial randomization is neither necessary nor sufficient for improving efficiency, identifying causal relationships, and discovering economically significant differences.[6]

Student's balanced, repeated, decision-theoretic, and power-oriented approach to the design and evaluation of experiments seems to be more valid in the ordinary sense of that word.[7] Balanced designs are deliberate or systematic arrangements of treatments and controls when there is one or more unobserved and non-random source of fluctuation in the output of interest (Student 1911, 1923)—such as when a differential fertility gradient cuts across a farm field, as it always does (Student (1911, 1938); Es, Gomes, Sellmann, and van Es (2007)), or when unobserved attitudes about schooling, gender, and race contribute to selection bias in the

---

[5] For discussion of Fisher's method and how it differs from the methods of Student, Egon Pearson, Harold Jeffreys, and other modern masters – the decision-theorists and Bayesian experimentalists not mentioned by Levitt and List - see Zellner (2004), Ziliak (2008), Ziliak and McCloskey (2008, chps. 1, 17-23), and McCloskey and Ziliak (2009).

[6] William S. Gosset (1876-1937) aka Student was a self-trained experimentalist and innovator of statistical methods who worked for Guinness's brewery his entire adult life: Apprentice Brewer (1899-1906); Head Experimental Brewer (1907-1935); Head of Statistics Department (1922-1935); Head Brewer, Park Royal location (1935-1937); Head Brewer, Dublin and Park Royal (1937) (see Ziliak (2008) for discussion).

[7] Student's legacy, explicitly acknowledged and not, is extensive: Pearson (1938, 1939, 1990), Neyman (1935, 1938), Jeffreys (1939 [1961]), Savage (1954, 1976), Kruskal (1980), Zellner and Rossi (1986), Heckman (1991), Press (2003), and Heckman and Vytlacil (2007), for example.

outcomes of social programs, as they often do (Heckman and Vytlacil (2007); Bruhn and McKenzie (2009)). Balancing means creation of symmetry in all of the important sources of error, random and systematic. Unfortunately, Levitt and List (2009, hereafter "L and L") do not discuss Student's methods and legacy of balanced designs. For example, the article does not mention the Student-Fisher debates of the 1920s and 1930s. [8] There are numerous other, more minor errors in Levitt and List (2009) that will not be discussed at length in this comment.[9]

2. Three theses

For ease of exposition each of the three main claims by Levitt and List (2009) is named by the thesis it represents. They are:

The first wave thesis. The claim is that the "first wave" (L and L, pp. 1-2) or "period" (p. 1) of field experiments in economic history can be traced to the "1920s and 1930s" (p. 1) in works by Ronald A. Fisher (p. 3) and Jerzy Neyman (p. 3-4). The authors single out Neyman and Fisher on the basis of, they say, their introduction of "randomization" (L and L abstract, pp. 1, 2, 3 *passim*) to the mathematical design of agricultural field experiments;

---

[8] See Student (1938), for example, and Ziliak and McCloskey (2008), chps. 20-22.

[9] For example, in the one paragraph Levitt and List (2009, p. 4) give to Gosset there are at least nine (9) errors:

> "Student'' (W.S. Gossett) was a statistician and chemist responsible for developing procedures for ensuring the similarity of batches of Guiness at the Guiness brewery. In this capacity, he developed the t-test (often denoted the ''Student's t-test'') as a technique to measure deviations of the sampled yeast content to the brewery's standard. However, because the brewery did not allow employees to publish their research, Gossett's work on the t-test appears under the name 'Student.'

The errors, with corrections, are: (1) the Gosset surname has one "t", as in one t-distribution (Concise Dictionary of National Biography 1961, p. 177); (2) Gosset trained at Oxford in chemistry but he never worked as a chemist. He was a brewer and businessman, from 1899 to 1937 (Ziliak 2008); (3) Gosset was studying small samples of barley, hops, and malt, in 1904, when he discovered the need for a small sample test: Gosset (1904); he was not studying "yeast content" (L and L, p. 4) when he discovered the trouble with small numbers; (4) In 1908 he was working on Student's *z*-test (Student 1908, pp. 13-19); the term "*t*-test" did not appear until 1925 or later, when Student (1925, pp. 105-106) changed the name of *z* to *t*; (5) Gosset's main job was mixing new beers and trying out new ingredients, by combining laboratory and field experiments (Ziliak 2008). His main job was not "ensuring the similarity of batches" though in Student (1927) he revealed techniques of industrial quality control he pioneered at Guinness beginning in 1904; (6) and (7) Guinness is spelled with two n's; (8) Gosset still has one "t"; (9) Contrary to the claim made by Levitt and List, Guinness employees were encouraged by the company to publish research; the conditions were that they publish under a pen name and avoid discussion of beer. In *Biometrika* alone, for example, there were numerous contributions by Guinness employees: 14 articles by Student alone and one or more each by his assistants, "Mathetes," "Sophister," and others (Ziliak and McCloskey 2008, pp. 213, 217).

The randomization thesis. Their second claim is, like the first, both historical and methodological in nature. Following the advice of "Fisher" (L and L, p. 3) and supposedly of "Neyman" (p. 3), List and Levitt claim that "randomization" of design (p. 4)—the blind and random assignment of treatments and controls to the experimental unit—is the "lynchpin" (p. 4)—the "foundation" (p. 1)—of any well-designed experiment.

Levitt and List go a step further with the randomization thesis, and assert that randomization is the only "valid" (L and L, p. 4) justification for use of Student's test of statistical significance (p. 4). "Significance" (p. 4)—a low level of Type I error—is in Levitt's and List's view the critical test.[10] And finally:

The third wave thesis. According to Levitt and List, the history of field experiments in economics can be divided into three "distinct" (pp. 1, 2, 15) waves or periods. The "first wave" is, to repeat, the agricultural field experiments of the 1920s and 1930s, wherein Fisher and Neyman are said to have "introduced" the randomization thesis.

The "second wave" (p. 2) is, they claim, represented by the government-sponsored social welfare experiments such as the SIME/DIME negative income tax experiments of the middle and later years of the 20th century (Zellner and Rossi 1986; Heckman and Smith 1995).

The "third wave" (p. 2) is described by Levitt and List as an original contribution to scientific method: economists doing randomized experiments with firms in the private sector of the economy (L and L, p. 15; see also: Herberich, Levitt and List 2009 and Harrison and List 2004). "We are currently in the third wave, which began in earnest roughly a decade ago, in the late 1990s" (L and L, p. 15). "This third wave has brought with it a much deeper and broader exploration of economic phenomena than was pursued in the earlier waves of field experimentation" (p. 15).[11]

The basis for their third claim is their belief that they are the first in history to apply artificial randomization to questions of "Industrial Organization" (L and L, p. 2). "Emerging from this third wave of field experimentation is an approach that we view as an important component of the future of natural field experiments: collaborating with outside private parties in an effort to learn about important economic phenomena" (L and L, p. 18).

This comment does not seek to reply to all of the claims made by Levitt and List (2009), only the most surprising ones. Their three main theses are examined below, in order of appearance.

3. The first wave thesis

> Our discussion focuses on three distinct periods of field experimentation that have influenced the economics literature. The first might well be thought of as the dawn of ''field'' experimentation: the work of Neyman and Fisher, who laid the experimental foundation in the 1920s and 1930s by conceptualizing randomization as an instrument to achieve identification via experimentation with agricultural plots (L and L, p. 1).

[10] Contrast the unanimous rejection of statistical significance by the Supreme Court of the United States, Matrixx Initiatives, Inc. v. Siracusano, No. 09-1156, March 22, 2011 (Ziliak (2011a)).
[11] Thus List and Shogren (1998) is an early contribution to the third wave, so conceived.

The first misstep is historical. If the goal is to identify early examples of field experiments in economics, the clock of history should start quite a bit earlier than they claim it does: consider, for example, *The Farmer's Letters* (1767), by Arthur Young and, more importantly, given the authors' statistical focus: Wood and Stratton (1910), Mercer and Hall (1911), and especially Student (1911, 1923).

For instance, Student's (1911) "Appendix" to the Mercer and Hall (1911) experiment was published 79 years before Splawa-Neyman (1923 [1990]) was translated into English, and had a far greater impact on the literature. Student's 1911 article—reprinted in *Student's Collected Papers* (1942, Pearson and Wishart, (Eds.))-is not mentioned by Levitt and List. Yet as Mercer and Hall (1911, p. 127) themselves said about their seminal field experiment published in the *Journal of Agricultural Science*: "We are indebted to 'Student,' by whose assistance and criticism we have been greatly aided in the whole of this discussion of our experimental results, for the working out of a method whereby the experimental error may be still further reduced when only a single comparison is desired, as for example between two varieties or two methods of manuring, by taking advantage of the correlation which exists between adjacent areas. This contribution [that is, Student's contribution of 1911] is set out in an Appendix."

Levitt's and List's omission of Student (1911) in their history is difficult to understand. He titled his 1911 Appendix "Note on a Method of Arranging Plots so as to Utilize a Given Area of Land to the Best Advantage in Testing Two Varieties" (see also: Student (1923); Gosset (1936); and Beaven (1947), pp. 238, 254-255).

In 1911 Student found (using Mercer's and Hall's mangolds data) that the standard deviations of mean yield differences were reduced as the experimental plot size was reduced. The reason has to do with a non-random, systematic, and confounding variable. The uncontrolled source of fluctuation spoiling random arrangements is, Student argued from 1911, the non-random occurrence of fertility gradients in the soil (see also: Student 1938; compare Pearson 1938). The diminishing marginal productivity of the soil as one travels from one side of the field to the other, and indeed from one block to another, must figure into the design of the experiment, Student found, or else an important source of crop growth will bias results.

Student made several advances in 1911. More than two decades before Fisher (1935) wrote about "the concepts of repetition, blocking, and randomization" (L and L, p. 3) Student found that the closer in space the competing varieties and/or treatments are sown together (through blocking or pairing—what Student (1911 [1942], p. 52) called "the principle of maximum contiguity") the more precise the standard deviations of mean yield differences between varieties and treatments.[12] In an admittedly crude way, Student suggested and proved in

[12] What is a "block"? Box, Hunter and Hunter (2005, p 92) explain that "A block is a portion of the experimental material (the two shoes of one boy, two seeds in the same pot) that is expected to be more homogenous than the aggregate (the shoes of all the boys, all the seeds not in the same pot). By confining comparisons to those within blocks (boys, girls), greater precision is usually obtained because the differences associated between the blocks are eliminated." Given the definition of a block, it is clear from the above discussion that Student (1911) has priority over Fisher (1925) and Fisher (1935). Fisher did not "introduce" (L and L, p. 3) blocking in the 1920s and 1930s though L and L join others asserting he did (for example, Bruhn and McKenzie (2009)).

his 1911 Appendix to the Mercer and Hall experiment that random layouts are inferior to deliberate balancing, and in four of the key variables relevant to farming and economic statistics: precision, efficiency, simplicity, and power to detect a large and real treatment and/or varietal difference when the difference is there to detect.

Said Student (1911 [1942], p. 49): "The authors [Mercer and Hall] have shown that to reduce the error as low as possible it is necessary to "scatter" the plots."

> I propose to deal with this point in the special case when a comparison is to be made between only two kinds of plots, let us say two varieties of the same kind of cereal.
>
> If we consider the causes of variation in the yield of a crop it seems that broadly speaking they are divisible into two kinds.
>
> The first are random, occurring at haphazard all over the field. Such would be attacks by birds, the incidence of weeds or the presence of lumps of manure. The second occurs with more regularity, increase from point to point or having centres from which they spread outwards; we may take as instances of this kind changes of soil, moist patches over springs or the presence of rabbit holes along a hedge.
>
> In any case a consideration of what has been said above will show that any "regular" cause of variation will tend to affect the yield of adjacent plots in a similar manner; if the yield of one plot is reduced by rabbits from a bury near by, the plot next it will hardly escape without injury, while one some distance away may be quite untouched and so forth. And the smaller the plots the more are causes of variation "regular"; for example, with large plots a thistly patch may easily occur wholly within a single plot leaving adjacent plots nearly or altogether clean, but with quite small plots one which is overgrown with thistles is almost sure to have neighbours also affected.
>
> Now if we are comparing two varieties *it is clearly of advantage to arrange the plots in such a way that the yields of both varieties shall be affected as far as possible by the same causes to as nearly as possible an equal extent.*
>
> To do this it is necessary, from what has been said above, to compare together plots which lie side by side and also to make the [side by side] plots as small as may be practicable and convenient.
>
> Student 1911 [1942], p. 49; emphasis added

He (p. 50) continued:

> Obviously nothing that we can do (supposed of course careful harvesting) can now alter the accuracy of the resulting comparison of yields, but we can easily make different estimates of the reliance which we can place on the figures.
>
> For example, the simplest way of treating the figures would be to take the yields of the plots of each variety and determine the standard deviation of each kind. *Then from published tables* [found in Student (1908), not cited by Levitt and List] *we can judge whether such a difference as we find between the total yields is likely to have arisen from chance.*
>
> An advance on this is to compare each plot with its neighbour and to determine the standard deviation of the differences between these pairs of adjacent plots.
>
> From what has been said above as to the occurrence of "regular" sources of error it will be seen that such differences as these will be to a much larger extent dependent on the variety, and to a lesser extent on errors, than if the mere aggregates are compared.

Student (p. 51) calculated the savings of land utilization to be expected by the farmer at various yields and levels of precision as measured by the standard deviation of mean yield difference. He found that "Roughly speaking one-twentieth acre plots of mangolds would require at least twice as much land as one-two-hundredth acre plots in order that we may place as much confidence in the result, while one-fiftieth acre plots of wheat would probably require more than twice as much as one-five-hundredth acre plots" (Student, p. 52)

He showed these impressive results in a table together with evidence that the standard deviations of comparison rise linearly or nearly so with increases in plot size. "Hence," he concluded, "it is clearly of advantage to use the smallest practicable size of plot, using chessboards and the principle of maximum contiguity" (Student, p. 52; compare Bruhn and McKenzie (2009) and Carson and others (2009)).

What Student did next—comparing the precision and efficiency of balanced versus random layouts— seems surprising in light of Levitt's and List's claims about randomization in field experiments:

> Also the advantage of comparing adjacent plots is apparent in these examples, since with [mangold] roots less than two-thirds of the land is required to give the same accuracy as random comparison and with the wheat less than half.
>
> Student 1911 [1942], p. 51.

Neglect of Student (1911) is unfortunate, as Student uses (and in some cases introduces) in the 1911 article the concepts of blocking, pairing, repetition, Student's *z*-table and test of significance, efficiency, precision, and balanced versus random designs of experiments.[13] Indeed, the origin of today's 'paired *t*-test' can be found in Student (1911). [14] Clearly field experiments

---

[13] Duflo, Glennerster, and Kremer (2007), like Levitt and List (2009), have credited Fisher with the concepts of replication and blocking. The misattribution of credit is an example of a Fisher bias in the literature, discussed at length by Ziliak and McCloskey (2008). See also: Pearson (1990), Kruskal (1980), and Savage (1971).

[14] In 1934 Fisher wrote a letter to Student requesting reprints of Student (1908), Student (1911), and Student (1923). Fisher asked about the intellectual history of taking paired differences. Student replied that pairing is old, dating back he half-jokingly said to "Old Noah" and the Ark:

> St. James's Gate
> Dublin
> 16 January 1934
>
> Professor R. A. Fisher, Sc.D., F.R.S.
> The Galton Laboratory,
> University College,
> London, W.C.1.
>
> Dear Fisher,
>
> I am sorry to say that I can only let you have off-prints of the last of your three requests. However, the first one is merely of historical interest; the second is of no real

in economics had made a major advance in 1911 though to date Student has not been properly credited for it.

Prior to Fisher's *Statistical Methods for Research Workers* (1925), Student was frequently credited.[15] For example, the "Hall" of Mercer and Hall (1911) was Sir A. Daniel Hall, the Director of Rothamsted Experimental Station in the era before Fisher—who joined the station in 1919 (Hall 1905). The "general conclusions" (Mercer and Hall, p. 127) of Mercer and Hall are important but it is Mercer's and Hall's conclusion number "4" which deserves emphasis, as it reveals a contribution that Student made to experimental thought and Rothamsted, 8 years before Fisher was hired and 14 years before Fisher developed his own theory of experimentation (Mercer and Hall, p. 127):

> (4) For practical purposes the authors [Mercer and Hall (1911)] recommend that in any field experiment each unit of comparison (variety, method of manuring, etc., according to the subject of the experiment) should be given five plots of one-fortieth of an acre, *systematically distributed within the experimental area.*

Informed by Student's novel calculations, Mercer and Hall (p. 127) reported:

> This [paired, balanced, or systematic design of the field] will reduce the experimental error to within two per cent. of the result, if the land is at all suited for experiment; it does not however eliminate variations due to unequal effects of different seasons upon the varieties or the action of the manures under experiment. Such variations can only be eliminated by continuing the experiment for several years. Similarly variations induced by the type of soil can only be ascertained by repeating the experiments on several soils.

Levitt and List go off track when they assert that "In 1919, Ronald Fisher was hired [at Rothamsted] to bring modern statistical methods to the vast experimental data collected [there]," which is perfectly true, but then add: "Fisher . . . soon realized that the experimental approach at Rothamsted was crude—without replication and with less than efficient treatments—thus he began in earnest to influence experimental design" (L and L, p. 3).

This would be a fascinating observation to make—and a real feather in Fisher's cap—if it did not deviate from the truth. First, the balanced design recommended by Student and accepted by Mercer and Hall was not "crude"; it was *more* efficient (Student proved in his tables) than random layouts. Second, Fisher did not begin in earnest in 1919 to "influence experimental design". Fisher did not begin to work on experimental design until late 1923 or 1924, and only *after* he got a letter from Student (discussed below) criticizing Fisher and Mackenzie (1923).

> interest to anyone but myself, and only that because I put the blame for what the Americans are pleased to call "Student's method", i.e., taking differences, fairly and squarely on Old Noah's shoulders.
>
> . . . .
>
> Yours very sincerely,
>
> W.S. Gosset

(Addendum to Gosset (1962), Vol. 1. In: Egon Pearson Papers, Box G9, University College London, Special Collections Library).

[15] For discussion, see Neyman (1938); Pearson (1938); Ziliak and McCloskey 2008, chps. 20-22.

Why, then, have Levitt and List credited Fisher in the 1920s and 1930s for originating field experiments in economics? The reason seems clear now: "Of course," they assert, "randomization was the lynchpin as the validity of tests of significance stems from randomization theory" (L and L, p. 4).

Elsewhere they write (L and L, p. 2): "the work of Fisher and Neyman in the 1920s and 1930s is worthwhile to . . . consider for two reasons. First [they claim again without proof] these experiments helped to answer important economic questions regarding agricultural productivity . . . Second, these studies are generally believed to be the first to conceptualize randomization as a key element of the experimental method."

To answer their main question—when in history did field experiments in economics begin? (p. 1)—the authors' surveyed history trying to find not the origins of field experiments in economics but rather the origins of artificial randomization in economics experiments (L and L, abstract, pp. 1, 2, 4; see also: J-PAL 2010). Yet randomization is not the purpose of an experiment in economics and in any case randomization does not constitute the peak of experimental method in economics (Heckman and Vytlacil (2007)). As Student (1923) was first to show, artificial randomization—however valuable it may be for testing and estimation when errors are independent and identically distributed (Student (1908))—is not as precise as balancing when observations and errors are correlated. I return to these fundamental issues in section 5.

4. Neyman sides with Student, preferring balanced designs

A little digging in the archives suggests that Neyman would reject the authors' claims, too. For example, in a *Journal of the Royal Statistical Society* obituary notice, Neyman (1938) gave highest honors to the subject of his notice: Student.[16]

Neyman considered Student's 1923 *Biometrika* article "On Testing Varieties of Cereals" to be the starting point of theory. (He did not know about Student (1911).) In the article celebrated by Neyman, Student took his 1911 insights and ambition to a new and higher level. He compared as no one had before the precision, power, and efficiency of balanced versus random designs of field experiments, in his study of Irish and English barley yield (compare Student (1938) and Pearson (1938)). The barley experiments were either designed or advised by Student who had worked since 1904 with the Irish Department of Agriculture in cooperation with dozens of farmers and maltsters scattered around Ireland and England (Gosset 1936). Said Neyman (1938, p. 228):

> As a third example of Student's pioneer work I shall quote his first paper[17] on agricultural experimentation, which should be rightly considered as a starting point of an extensive theory which is now well known. There were several further papers by Student on the subject, including his last [Student 1938], now in print in *Biometrika*, which deals with the difficult problem of the advantages and disadvantages of systematically balanced layouts.

[16] Contrast Neyman's mainly negative assessment of Fisher's approach: Neyman (1961).

[17] In fact, Student (1911) was Gosset's first published article on the design and evaluation of agricultural field experiments. See Student (1942) for the full collection of Student's published articles.

This was not the first time that Neyman had sided with Student. At a 1935 meeting of the Royal Statistical Society, Neyman (1935, p. 109) said: "Owing to the work of R. A. Fisher, Student and their followers, it is hardly possible to add anything essential to the present knowledge concerning local experiments. There remain only details," he said, "which perhaps require some more attention than has been given to them before."

He did not mean to put Student and Fisher on equal footing. Neyman (1935, p. 173) clarified his point of view in reply to criticisms made by Fisher at that same meeting of the Royal Society: "I am considering problems which are important from the point of view of agriculture," he carefully emphasized. "My point of view," he told Fisher, "is shared by other writers; for instance "Student," in his classical memoir published in Vol. XV of Biometrika" (Student 1923).

Levitt and List claim otherwise, arguing that Neyman "laid the experimental foundation" (p. 1) with randomization theory (L and L, pp. 3-4): "Viewing Neyman's body of work [meaning the one article they cite by Splawa-Neyman (1923)], we find it clear that early on he understood deeply the role of repeated random sampling and that a necessary condition for probabilistic inference is randomization" (L and L, p. 3).

Again, this would be a wonderful finding if it were close to the truth. It is not. First, the Splawa-Neyman (1923) article did not affect the development of first wave experiments. The article was not translated into English from the original Polish until 1990—almost ten years after Neyman died and decades after the "first wave" of agricultural experiments commenced (see Splawa-Neyman [1923], translated and edited by D. M. Dabrowska and T.P. Speed, 1990).

Second, as Neyman himself explains, Neyman did not advocate use of artificial randomization for designing experiments: he sided with Student's balanced approach. In Neyman's view, randomization was neither necessary nor sufficient for the "extensive theory" (Neyman 1938, p. 228) which Student (1923) developed to solve "the difficult problem" (Neyman, p. 228) of design (see also: Reid 1982, pp. 44).

Neyman did not claim that random layouts have no value; rather, he, like Student before him, was forced by logic and fact to concede the economic and statistical advantages afforded by Student's balanced layouts. Neyman admitted that balancing involves some sacrifice of the normal assumptions. He said in his reply to Fisher (Neyman 1935, p. 179):

> As to the Half-Drill-Strip method [a balanced layout sometimes called "ABBA" (Student 1936, p. 197)], I must agree that from an orthodox statistical view-point it is not quite correct - that is to say that the estimate of error variance [in the balanced layout] is not quite correct.[18]

"In a publication already referred to," Neyman said, "I tried to improve the method in this respect" (compare Pearson 1938 and Neyman and Pearson 1938). "But, then, from the same orthodox viewpoint, which of the other methods is absolutely correct in all its details? The important question," Neyman said, "is which of the inaccuracies in which method has a smaller numerical value. This requires special investigation. But my personal feeling is that it would be altogether wrong to attach to the Half-Drill-Strip method [Student's preferred method] less importance than to any other method in frequent use, this especially because of the 20 replications which it is usual to associate with it" (Neyman 1935, p. 179).

[18] Student (1923, 1936, 1938) was the first to emphasize this minor weakness of balancing.

Given these facts about Student and Neyman, it is not clear what Levitt and List hoped to achieve with their "glimpse" (L and L 2009, p. 15) at randomization in history.[19]

4.1 Fisher did not understand randomization in 1923

According to Levitt and List (2009, p. 4) "Fisher and McKenzie (1923)" is the second classic article to use randomization in the design of a field experiment. This is a remarkable achievement given that randomization does not appear even once—randomization is neither used nor mentioned—in the article by Fisher and Mackenzie.

Here is how Levitt and List (2009, p. 3) describe Fisher's alleged 1923 contribution to randomization in field experiments: "Fisher's fundamental contributions were showcased in agricultural field experiments. In his 1923 work with McKenzie [sic], Fisher introduced . . . randomization" (Fisher and McKenzie, 1923)" (Levitt and List, p. 4). But that is not so; what they are claiming is not true. In fact it is precisely the absence of careful planning which made the 1923 Fisher and Mackenzie experiment infamous—famous in the bad sense—eliciting negative comments from Student, Yates, Cochrane, and others.

The fact of the matter is that in 1923 Fisher had not given much if any thought to the statistical design of experiments. Cochrane (1989, p. 18) notes that:

> Fisher does not comment [in Fisher and Mackenzie (1923)] on the absence of randomization or on the chessboard design. Apparently in 1923 he had not begun to think about the conditions necessary for an experiment to supply an unbiased estimate of error.
>
> Cochrane, p. 18, in Fienberg, Hinckley, eds., 1989

Yates (1964, pp. 310-311) goes further. Like Cochrane, Yates observes that in 1923 Fisher did not possess any theory of experiments, random or balanced. Says Yates (pp. 310-311) of Fisher's and MacKenzie's 1923 manure experiment:

> Twelve varieties of potatoes were grown with two types of potash (sulphate and chloride) and also without potash. Half the experiment also received farmyard manure. There were three replicates of each variety on each half, each varietal plot being split into three in a systematic manner for the potash treatments. The actual layout (Fig. 1) [by Fisher and Mackenzie] illustrates [Yates said] how little attention was given to matters of layout at that time.[20] It is indeed difficult to see how the arrangement of the varietal plots [designed by Fisher and Mackenzie] was arrived at.

---

[19] It is interesting to note that Neyman, a great mathematician, did not think that mathematics is the best guide when designing experiments with an economic motive. "I will now discuss the design of a field experiment involving plots," he began his article of 1923. "I should emphasize," he said, "that this is a task for an agricultural person [such as Student] however, because mathematics operates only with general designs" (Splawa-Neyman 1923 [1990], p. 465; translated and edited by D. M. Dabrowska and T.P. Speed, 1990).

[20] But compare Student (1911, 1923), two seminal articles conspicuously omitted by Yates.

Fisher's approach in 1923 was neither randomized nor balanced: "the arrangements for the farmyard manure and no farmyard manure blocks are almost but not quite identical, and some varieties tend to be segregated in the upper part and others in the lower part of the experiment" (Yates, pp. 310-311). "Consequently," wrote Yates, "no satisfactory estimate of error for varietal comparisons can be made [in the Fisher and Mackenzie experiment]. . . .To obtain a reasonable estimate of error for these interactions," he said, "the fact that the varietal plots were split for the potash treatments should have been taken into account. This was not done in the original analysis" (Yates 1964, pp. 310-311). Levitt and List lead readers to believe otherwise.

Yates continued (Yates 1964, pp. 311-312): "The principle of randomisation was first expounded in [Fisher's textbook] *Statistical Methods for Research Workers* [1925]"—not in Fisher and MacKenize (1923). In other words, an article that Levitt and List claim for the canon in the history of randomization neither mentions nor uses randomization.

It is interesting to pursue the social ramifications of Fisher's spoiled experiment. Fisher asked Student for his opinion on Fisher and MacKenzie (1923). Student was glad to opine, for he had been a contributor to the *Journal of Agricultural Science* since 1911, and had already noticed the 1923 article. He told Fisher in a letter of July 1923 about the article which Levitt and List consider foundational: "I have come across the July J. A. S. [*Journal of Agricultural Science*] and read your paper," Student said, and

> I fear that some people may be misled into thinking that because you have found no [statistically] significant difference in the response of different varieties to manures that there isn't any. The experiment seems to me to be quite badly planned, you should give them a hand in it; you probably do now."[21]

Ouch. That must have hurt Fisher, whose career was just getting started. Still, to his credit, Fisher replied to Student's helpful if embarrassing letter, asking Student what he, Student, would do differently.[22] Student replied in a letter of July 30, 1923: "How would I have designed the [Fisher and Mackenzie] exp.[eriment]? Well at the risk of giving you too many 'glimpses of the obvious'," Student—the Head Experimental Brewer of Guinness, inventor of Student's t, and a pioneering lab and field experimentalist with by then two decades of experience told the novice at Rothamsted—"I will expand on the subject: you have brought it on yourself!", Student told Fisher in 1923. "The principles of large scale experiments are four":

> There must be essential similarity to ordinary practice . . . . Experiments must be so arranged as to obtain the maximum possible correlation [*not* the maximum possible statistical significance] between figures which are to be compared. . . . Repetitions should be so arranged as to have the minimum possible correlation between repetitions . . . There should be economy of effort [maximizing net "pecuniary advantage" in Gosset's (1905a) sense, as discussed by Ziliak (2008)].[23]

Fisher did not take to Student's balanced and economic approach, a point I take up immediately below. In sum, Levitt and List tell an erroneous history of randomization. Notably, Fisher and MacKenzie (1923) – a famously spoiled experiment - does not mention randomization let alone use it. Their selection of articles for the canon is not credible.

---

[21] Letter #29, July 25, 1923, in Gosset 1962.
[22] Letter #s 20, 23, in Gosset 1962.
[23] Letter #s 29, 30, in Gosset 1962; Gosset 1923, in E. S. Pearson 1990 [posthumous], p. 58.

4.2 Fisher and the false sociology of randomization, 1923 to 1925

Student's second letter to Fisher must have put in a bee in his bonnet.[24] Regardless, between August 1923 and October 1924, Fisher for some reason became a radical proponent of artificial randomization in the design of experiments—the very design which Student (1911, 1923) disproved. As Yates (1964) notes, Fisher went a step further in 1925 and – against Student's advice - presented randomization as a first principle of design.

Was Fisher rebelling against his mentor, the older and more experienced Student (Gosset 1962)? Perhaps. As Cochrane and Yates note, Fisher did not discuss randomization of design until *Statistical Methods for Research Workers* (1925)—only two years after Fisher and MacKenzie (1923) and the letter from Student explaining to Fisher the basic principles of experimental design. In late 1923 and throughout 1924 Fisher continued to seek Student's assistance. For example, they exchanged many letters about Student's t table which Student prepared at Fisher's request for publication in *Statistical Methods for Research Workers* (1925, "The table of t"). Interestingly, Fisher did not reply to Student's letter of July 1923, the one claiming (correctly as it turns out) that Fisher in 1923 had (1) no theory of experimental design and (2) misused Student's test of statistical significance.

Fisher's 1925 declaration of a randomization "principle" was not warranted by the evidence accrued by Student. In October 1924 Fisher wrote to Student asking for pre-publication comments on Fisher's now-classic textbook. Seeing a set of pre-publication page proofs (which Fisher had sent to him) with unusually few corrections to the experimental sections, Student suggested to Fisher that the lack of corrections may be "possibly because of his [Fisher] understanding less of the matter."[25] "I don't agree with your controlled randomness," Student explained in another important letter to Fisher, in October 1924. "You would want a large lunatic asylum for the operators who are apt to make mistakes enough even at present," Student said of his so-called randomization "principle". "If you say anything about Student in your preface you should I think make a note of his disagreement with the practical part of the thing." Significantly, Fisher did not say anything of the kind in the preface about Student (1911, 1923); he did not mention the opposition and priority of the anonymous brewer whose innovations by 1923 clearly included comparison of the advantages and disadvantages of balanced versus random designs of experiments: Student did this work two years before Fisher first wrote about "randomized blocks".

---

[24] In Fisher (1933) it seems to be more than chance which led him to omit mention of Student (1911, 1923, 1926) and Student's comparisons of random with balanced designs of experiments. He did not mention Student's disagreement, though mention was requested by Student in the 1924 letter. Significantly, in addition, Fisher's 1933 history of experimental statistics adopts a mocking tone whenever the author is speaking about his predecessors—the foolish "pride" (Fisher 1933, p. 46), he said they took, in "High correlations" (ibid, p. 46) prior to Fisher (1925). If Student ever lost patience when dealing with Fisher's difficult personality it was in Student's last published article, "Comparison Between Balanced and Random Arrangements of Field Plots" (Student 1938 [posthumous]); see also Ziliak and McCloskey 2008, chps. 20-24).

[25] Quoted in Pearson 1990, p. 59

The article by Levitt and List, however flawed, does raise an interesting question for the history of science: who introduced the idea of artificial randomization into statistically-based experiments in economics?

Normally people answer as Cochrane, Yates, and now Duflo and Levitt and List do, crediting Fisher. Yates (1964), who for many years worked side by side with Fisher, and co-authored books and papers with him, acted as if he was certain that randomized blocks were Fisher's original inspiration but then he admitted in the next sentence that he doesn't know of evidence that would say much one way or another about the accuracy of his claim.

It seems at least as valid, maybe more valid, to assert that Student (1908, 1911) and especially Student (1923) deserve the credit—the credit for first introducing—and then also for rejecting—use of artificial randomization in the design of field experiments in economics.

The reason that artificial randomization was "in the air" (Levitt and List, p. 4) is because Student had helped to put it there years before, in a seminal Monte Carlo simulation he describes in his 1908 *Biometrika* article "On the Probable Error of a Mean":[26]

> Before I had succeeded in solving my problem analytically [Student's *z* distribution], I had endeavoured to do so empirically. The material used was a correlation table containing the height and left middle finger measurements of 3000 criminals, from a paper by W. R. MacDonell (*Biometrika*, Vol. I., p. 219). The measurements were written out on 3000 pieces of cardboard, which were then very thoroughly shuffled and drawn at random. As each card was drawn its numbers were written down in a book, which thus contains the measurements of 3000 criminals in a random order. Finally, each consecutive set of 4 was taken as a sample—750 in all—and the mean, standard deviation, and correlation of each sample determined. The difference between the mean of each sample and the mean of the population was then divided by the standard deviation of the sample, giving us the *z* of Section III
>
> Student 1908, p. 13

And the reason for Student's 1923 article was precisely to demonstrate the economic and statistical advantages and disadvantages of balanced versus random layouts of competing varieties in a small number of repeated experiments. Levitt and List cite Student (1923) but they fail to mention its purpose and main finding, which is against randomization and in favor of balancing. What they could have noted in their article but did not is that Student's pioneering rejection of randomization came two years before Fisher (1925) first wrote about randomization.

Ironically, Student's approach was economistic through and through, though neglected by Levitt, List, and other experimentalists in the current generation. Student (1923, p. 281) compared 193 plots on 18 farms for 8 different varieties of barley grown in the barley regions of Ireland:

---

[26] I am not claiming that Student (1908) is the first use of artificial randomization in a formal statistical context. That honor belongs to others, such as Peirce and Jastrow (1885) (Stigler 1986, p. 253). I am only claiming that Student's (1908, p. 13) simulation of the small sample *z*-distribution for a profit-seeking firm (Guinness) is a seminal use of artificial randomization in experimental economics (see also: Pearson (1990)).

> *If* now the plots had been *randomly* placed [rather than balanced in chessboard fashion], the variance of a comparison between two of the races would been approximately 228, and about 25 times as much ground would have been required to reduce the standard error of a comparison to 1%.

From "28 variances" (Student 1923, p. 282) calculated using Fisher's ANOVA structure, the Guinness scientist concluded:

> In other words, we have gained by chessboarding to the extent that we are as accurate as if we had devoted twice the area to plots randomly arranged.

Going further, Student (1923, p. 285) introduced to Biometrika an improvement over chessboard and random, innovated by him and Beaven, "The Half-Drill Strip Method", otherwise known as "ABBA" (Student 1938, pp. 364-378). Student and his collaborator E.S. Beaven (1947) found in repeated experimentation that when testing a new variety or treatment against the standard variety or treatment - crop or fertilizer, say - the half-drill strip method was more accurate and cost efficient than both chessboards and random:

> We now proceed to the most accurate method yet devised for field trials.
>
> Student 1923, p. 285

5. Student and the higher power of ABBA: Against the randomization thesis

> Is there some design
> In these deep random raindrops
> Drying in the dust?
>
> Wright 1998, p. 34

What made the half-drill strip or ABBA method so effective, compared to random and the other balanced designs? In general if a variable in the environment exhibits a systematic spatial or temporal correlation with respect to the experimental output of interest then the allocation chosen by pure randomization is likely to bias results; coefficients will be wrong, inferences will be incorrect, and power – not to mention guesses about profitability - inadequate. This is as true of medical phenomena - Altman, Rothman, and others have found - as it is of social policy and development economics (Heckman and Vytlacil 2007; Bruhn and McKenzie 2009; Ziliak 2010).

For instance, in barley yield trials, a significance test based on artificial randomization does not control for a major source of error, Student showed – differential fertility of soil - and gives in general less valid results compared to balanced designs. Early examples of balanced designs in crop yield trials are chessboard and Knight's move.

Take the Knight's move, for example, balancing an 8x8 chessboard-type design. Suppose a brewer is comparing yield of 8 different varieties of barley in the field—as Student (1923) was —the 8 varieties matching the 8 rows and columns an actual chessboard. How shall the seed be planted? Knight's move says to balance the distribution of real sources of error (that is the diminishing marginal productivity of the land) by allocating seeds of a unique variety as

one would a Knight's piece in chess – plant Variety A in a block that is two up and one over from a block occupied by A; Variety B, again, like the Knight's move in chess, should be assigned to a block that is one down and two over from a block occupied by one of its own kind, and so forth, for each variety and permutation of the experiment, given the chosen n x k design and the number of experiments (farms) in the series.

Take the simplest case, comparing yields of two different barleys, barley A (the standard variety) and barley B the new.

In the 8x8 chessboard layout the experimental field has n= 64 blocks in which one may randomly or deliberately grow variety A or variety B. (In a more complicated strategy, blocks may be further subdivided, such that individual blocks can grow seeds from A and B. We will stick to the simple event that each block gets a unique "treatment".) A random assignment of A and B to blocks may produce, for example, the following pattern:

A A A A A A B B
A A A A A A B A (i)
. . . . etc.
⟶ Direction of increase in soil fertility (higher yielding soil)

Another random draw may produce blocks of this sort:

B B B B B A A B
B B B B B A A A (ii)
. . . etc.
⟶ Direction of increase in soil fertility (higher yielding soil)

How precise are the estimated differences in average yields, A-B, or B-A, if fertility on the left side of the field is systematically lower than fertility on the right? Layouts such as (i) and (ii)—though random—produce biased mean squared errors and parameter estimates with respect to a major source of fluctuation—differential soil fertility. In example (i) the As are bunched up and growing in the very worst soil; thus the yield of the Bs will be artificially high, and the real treatment difference, A-B, will be undetermined.

Student found again and again that deliberate balancing—though adding to the "apparent" (Student 1938 , pp. 364-372) error, that is, to Type I error in ANOVA terms, actually *reduces* the real error of the experiment—minimizing Type 2 error and errors from fixed effects, such as non-random soil heterogeneity.

Examples (i) and (ii) suggest that whenever there is a systematically variant fertility slope (or other temporal or secular source of local and fixed effect) which cannot be artificially randomized, the systematic source of fluctuation cannot be ignored without cost: differences in yield will be correlated by local and adjacent fertility slopes—ground effects—any temporal or spatial or real difference which can't be randomized. Random layouts analyzed with Student's test of significance will yield on average more biased differences, A-B and B-A, and less ability to detect a true difference when the difference is large.

By 1923 Gosset's solution for dealing with systematic sources of variation between A and B became (grammarians have to admit) perfectly balanced: he dubbed his balanced design of choice, "ABBA" (Student 1938, pp. 364-378). The ABBA layout is:

A B B A A B B A (iii)
A B B A A B B A
A B B A A B B A
. . . etc.

One virtue of the ABBA design is that it minimizes bias caused by differential soil fertility. Given the built-in symmetry of ABBA, no matter what the trajectory or magnitude of differential fertility gradients, A's and B's are equally likely to be grown on good and bad soil. Random throws of seed do not have this virtue, biasing mean yield differences, A-B.

Yet ABBA brings additional statistical and economic advantages, too. On the supply side, with ABBA the ease and cost of sowing and harvesting and calculating basic statistics on yield is plot-wise and block-wise reduced. Compare the rows and columns of ABBA with the random rows and columns in (i) and (ii) above and it's easy to appreciate Student's sensitivity to supply side economic conditions.

With ABBA there is no need for chaotic tractor driving while planting seed in blocks randomly dispersed; and thus with ABBA there is a lot less measurement error and loss of material at harvest and counting time (see Beaven 1947 for exact details of his and Student's procedure). Imagine harvesting and counting up the mean difference in yield of strip A minus strip B, block by block, in the ABBA field versus the randomized and one can appreciate further still the efficiency of Student's balanced solution. As Student told Fisher in the letter of 1923, "There must be essential similarity to ordinary [in this case, farming] practice" (Pearson (1938, pp. 163-164) shows how to adjust Student's test of significance to accommodate the ABBA structure.) After all, "[t]he randomized treatment pattern is sometimes extremely difficult to apply with ordinary agricultural implements, and he [Student] knew from a wide correspondence how often experimenters were troubled or discouraged by the statement that without randomization, conclusions were invalid" (Pearson 1938, p. 177).

Fisher, for his part, rejected Student's ABBA and other balanced designs (see, for example, Fisher and Yates (1938), which fails to mention Student's methods). Student's (1938, p. 366) last article – which he worked on during the final months and days of his life and until the day he died – said to Fisher:

> It is of course perfectly true that in the long run, taking all possible arrangements, exactly as many misleading conclusions will be drawn as are allowed for in the tables [Student's tables], and anyone prepared to spend a blameless life in repeating an experiment would doubtless confirm this; nevertheless it would be pedantic to continue with an arrangement of plots known before hand to be likely to lead to a misleading conclusion. . . .
>
> In short, there is a dilemma—either you must occasionally make experiments which you know beforehand are likely to give misleading results or you must give up the strict applicability of the tables; assuming the latter choice, why not avoid as many misleading results as possible by balancing the arrangements? . . . To sum up, lack of randomness may be a source of serious blunders to careless or ignorant experimenters,

but when, as is usual, there is a fertility slope, *balanced arrangements tend to give mean values of higher precision compared with artificial arrangements*

Student 1938, p. 366.

What about variance? What affect does balancing have on variance and thus on the level of statistical significance?

"The consequence is that balanced arrangements more often fail to describe small departures from the 'null' hypothesis as significant than do random, though they make up for this by ascribing significance more often when the differences are large" (Student 1938, p. 367).

6. Illustration of ABBA and the Higher Power to Detect Large Treatment Effects

The intuition behind the higher power of ABBA[27] and other balanced designs to detect a large and real treatment difference was given by Student in 1911 (compare Beaven 1947, pp. 273-5). "Now if we are comparing two varieties it is clearly of advantage to arrange the plots in such a way that the yields of both varieties shall be affected as far as possible by the same causes to as nearly as possible an equal extent" (Student 1911, p. 128). He called this part of his theory, to repeat, the principle of "maximum contiguity" (Student 1923 [1942], p. 95)—a principle he put to work again and again, such as when he illustrated the higher precision and lower cost associated with a small-sample study of biological twins, to determine the growth trajectory of children fed with pasteurized milk, unpasteurized milk, and no milk at all, in "The Lanarkshire Milk Experiment" (Student 1931a, p. 405).[28]

---

[27] Student's and Beaven's ABBA design is, formally speaking, *chiasmus*—one of the most powerful and influential design patterns in the history of language, music, religion, and science. What is chiasmus beyond the symmetric Greek symbol for *chi*, X, from which the term derives? Lanham (1991, p. 33) defines chiasmus as "The ABBA pattern of mirror inversion". Unaware of Student's ABBA, the classical rhetorician explains: "Chiasmus seems to set up a natural internal dynamic that draws the parts closer together . . . The ABBA form," he notes, "seems to exhaust the possibilities of argument, as when Samuel Johnson destroyed an aspiring author with, 'Your manuscript is both good and original; but the part that is good is not original, and the part that is original is not good'" (p. 33). Good, original, original, good: the ABBA layout. James Joyce, another famous Dubliner in Student's day, wrote chiasmus in his novella *Portrait of the Artist as a Young Man*. Other examples of chiasmus are by John F. Kennedy ("Ask not what your country can do for you; ask what you can do for your country") and by Matthew 23:11-12 ("Whoever exalts himself will be humbled, and whoever humbles himself will be exalted"). In science, supply and demand and the double helix are two notable examples of chiasmus.

[28] Student (1931a, p. 405) estimated that "50 pairs" of [identical twins] would give more reliable results than the 20,000" child sample, neither balanced nor random, actually studied in the experiment funded by the Scotland Department of Health. "[I]t would be possible to obtain much greater certainty" in the measured difference of growth in height and weight of children drinking raw versus pasteurized milk "at an expenditure of perhaps 1-2% of the money and less than 5% of the trouble." Karlan and List (2007, p. 1777) could have revealed more about the economics of charitable giving—for less—using a variant of Student's method. Instead the *AER* article studied n=50,083 primarily white, male, pro-Al Gore donors to public radio, not balanced.

The power of balanced designs to detect real differences can be seen if one imagines doing as Student did, trying to maximize the correlation of adjacently growing varieties and/or treatments, the As and Bs. He then measured the difference of yields of strips (A-B and B-A) which he situated as closely as practicable from the farming, botanical, and price points of view. At harvest time he noted that the tractor can easily collect the rows of As and Bs, providing the farmer has used ABBA to layout the field; and it is easy to see, other things equal, that random planting is more costly at harvest time: compared to the neat rows and strips of ABBA, random layouts require far more labor and driving around in the tractor, if possible after random sowing.

In "Some Aspects of the Problem of Randomization: II. An Illustration of Student's Inquiry Into the Effect of "Balancing" in Agricultural Experiment," Egon S. Pearson (1938) – another skeptic not mentioned by the authors - clarified Student's theory of ABBA.[29] Said Pearson (1938, p. 177):

> In co-operative experiments undertaken at a number of centres, in which as he [that is Gosset aka Student] emphasized he was chiefly interested, it is of primary concern to study the difference between two (or more) "treatments" under the varying conditions existing in a number of localities.

For treatments and/or varieties A and B, Student's idea is to estimate from the ABBA experiment:

$$x_A = m_A + \delta_A$$

and

$$x_B = m_B + \delta_B$$

and thus: $x_A - x_B = (m_A - m_B) + \delta_A - \delta_B = (m_A - m_B) + \Delta_{AB}$ (iv)

(Pearson 1938, pp. 163-164) where x_i is the yield from the *i*th block or plot, m_i is the yield in the *i*th block or plot to which a treatment has been applied (an unchanging value no matter the treatment) and δ_i is the real treatment in the block or plot.

Students of Heckman, Friedman, and Zellner, for example, will not be surprised by what follows from Student's and Pearson's set up, which strives to achieve real error minimization. The comparative advantage of Student's and Pearson's ABBA design in repeated trials is: (1) ABBA enables explicit control of the *m's*—the difference in growing conditions or other fixed factor whose influence you are trying to minimize, and (2) ABBA enables more control of the variance of Student's $\Delta_{AB's}$ –the real treatment effects (or causes if you will) on yield, within and between farm fields.

> It has been said from an experiment conducted by this method no valid conclusion can be drawn, but even if this were so, it would not affect a series of such experiments.[30] Each

[29] Box 10, brown folder, Egon Pearson Papers, University College London, Special Collections Library.

[30] Beaven (1947, p. 293) reported after 50 years of experimentation on barley using his and Student's methods that selection of a new cereal takes "about ten years" (p. 293) of repeated and balanced experimentation. By the early 1920s three different varieties of barley, selected and proved by Beaven and Student, were grown on "well over five million acres of land" (Beaven, xiv). When Beaven died (in 1941) Guinness acquired ownership and continued to produce at the

> is independent of all the others, and it is not necessary to randomize a series which is already random, for, as Lincoln said, "you can't unscramble an egg". Hence, the tendency of deliberate randomizing is to increase the error.
>
> Gosset 1936, p. 118.

Using a simple workhorse formula, Student showed that the ABBA layout—arranging varieties or treatments A and B close together in strips—reduces the standard deviation of yield differences by maximizing ρ—the correlation between yields of the competing treatments and/or varieties, A and B. The formula he used as the basis for measuring the variance of mean differences, A-B, he got in 1905 from Karl Pearson, during a July visit to Pearson's summer house:

$$\sigma^2_{A-B} = \sigma^2_{A} + \sigma_{B}^2 - 2\rho_{AB}\sigma_A\sigma_B \qquad (v)$$

where σ^2 is variance and ρ is the Pearson correlation coefficient (Pearson quoted by Gosset 1905a, 1905b, Guinness Archives; see also: E.S. Pearson 1939, p. 212).[31]

Given the systematic way that the sun, wind, water, and other environmental features—such as rabbit holes and fertility gradients—affect growth potential in a given block of any field, the spatial closeness and symmetry of ABBA maximizes the size of ρ—exactly what the analyst wants when high power, efficiency, and equal balance of confounding errors are goals.

The higher the correlation ρ between yields A and B the lower is the variance of their differences A-B and B-A. Thus compared to random the ABBA design gets more statistically significant results when the differences between A and B are truthfully large—the power to detect is high when the effect size is large—exactly what the firm – such as a large scale brewery - wants when precision and profit are goals.

Fisher's randomization—and the new field experiments—ignore the fundamental importance of the correlation coefficient, ρ; assuming independent and identically distributed

---

famous farm and malt house in Warminster, UK (http://www.warminster-malt.co.uk/history.php). Contrast the new field experiments in economics, neither repeated nor balanced yet full of advice for going concerns (Herberich, Levitt, and List (2009), Levitt and List (2009), Banerjee and Duflo (2011), and by now hundreds of others).

[31] Though unaware of the Student-Pearson model for estimating the variance of real treatment effects in yield trials, Friedman (1953) used the same basic set-up to simulate the effect of a counter-cyclical expenditure (spent by a central government in search of full-employment) on the variance of aggregate income. "Let *X(t)* represent income at time *t* in the absence of the specified full-employment policy. The full-employment policy," Friedman said, "may be regarded as having effects that add to or subtract from income." Searching for something like Student's real treatment effect, $\delta_A - \delta_B$, Friedman continued: "Let *Y(t)* represent the amount so added or subtracted from *X(t)*, so that: *Z(t) = X(t) + Y(t)* represents income at time *t*. . . . What is the optimum size of σ^2_y?" Friedman asked. "By a well-known statistical theorem $\sigma^2_z = \sigma^2_x + \sigma^2_y + 2r_{xy}\sigma_x\sigma_y$ where r_{xy} is the correlation coefficient between X and Y" (Friedman 1953, pp. 122-123). Other things equal, the effect of the "stabilizing" action depends in large part on the magnitude of the correlation coefficient, just as Friedman and others after Student and Pearson have demonstrated in a vast body of literature ignored by Levitt and List (2009).

observations in imaginary replications, artificial randomization seeks only to minimize $\sigma^2_{\ A}$ and $\sigma_{\ B}{}^2$. Yet plot by plot as Student (1923, p. 273) said:

> The art of designing all experiments lies even more in arranging matters so that ρ [the correlation coefficient] is as large as possible than in reducing σ^2_x and σ^2_y [the variance].
>
> The peculiar difficulties of the problem lie in the fact that the soil in which the experiments are carried out is nowhere really uniform; however little it may vary from eye to eye, it is found to vary not only from acre to acre but from yard to yard, and even from inch to inch. This variation is anything but random [Student observes], so the ordinary formulae for combining errors of observation which are based on randomness are even less applicable than usual.

Thus it is quite misleading when Levitt and List (2009, p. 4) assert that "Gossett understood randomization and its importance to good experimental design and proper statistical inference."

When estimating how Δ_{AB} – the real treatment difference—varies from one set of conditions to another (for example from one farm to another) one is free to assume the validity of Student's table of *t* and test of significance. Randomness—not randomization—is all that one needs to justify use of Student's table, Student persuasively noted in 1908 (Student 1908a, pp. 1-2) and repeated in Student (1938).

In *Theory of Probability* (1961), "§4.9 Artificial Randomization," the great Bayesian experimentalist Harold Jeffreys (not mentioned by Levitt and List) agrees with Student. When fertility contours are present (and uniformity trials showed that they always were) "there is an appreciable chance that [the differences in soil] may lead to differences that would be wrongly interpreted as varietal [as relating to the barley rather than to the fixed features of the soil; in medicine think of the pill and the different abilities of hearts]" (Jeffreys 1961, p. 242). "Fisher proceeds . . . to *make it* into a random error" (p. 243; italics in original).[32] But:

> Here is the first principle [Jeffreys said]: we must not try to randomize a systematic effect that is known to be considerable in relation to what we are trying to find. . . The [balanced] method of analysis deliberately sacrifices some accuracy in estimation for the sake of convenience in analysis. The question is whether this loss is enough to matter, and we are considering again the efficiency of an estimate. But this must be considered in relation to the purpose of the experiment in the first place.
>
> Jeffreys 1961, p. 243

Thus a well-designed field experiment in economics strives for efficiency, and for the power to detect a minimally important difference, with a low real error. Fisher-randomization and significance, measured by the p-value, does not. Said Jeffreys (1961, p. 244), again, citing Student (1923, 1938) as the source of his ideas:[33]

> There will in general be varietal differences; we have to decide whether they are large enough to interest a farmer, who would not go to the expense of changing his methods

[32] The "it" is, in this case, the non-random distribution of soil productivity.

[33] Jeffreys considered Student's methods to be in basic agreement with his own Bayesian approach (Jeffreys 1961, pp. 379, 393, 369-400). Ziliak (2008) and Ziliak and McCloskey (2008) describe Student's early endorsement of Bayes's theorem in practical work.

> unless there was a fairly substantial gain in prospect. There is, therefore, a minimum difference that is worth asserting.
>
> Jeffreys 1961, p. 243

And to detect a minimum important difference, Student (1938) discovered in his last article - and Pearson's (1938, p. 177) simulations later confirmed - "a definite advantage that seemed to be gained from balancing". Exactly as Student expected, Pearson found that when treatment and/or varietal differences grow large, the power curves of balanced and random designs cross, lending advantage of detection to balanced designs (see Figure 1; see also: Bruhn and McKenzie (2009)).

Student put benefit and cost at the center of experimental design and evaluation. Indeed, as Student wrote in an important letter of 1905 to Egon's father, Karl Pearson:

> When I first reported on the subject [of "The Application of the 'Law of Error' to the Work of the Brewery" (Gosset, 1904)], I thought that perhaps there might be some degree of probability which is conventionally treated as sufficient in such work as ours and I advised that some outside authority in mathematics [such as Karl Pearson] should be consulted as to what certainty is required to aim at in large scale work. However it would appear that in such work as ours the degree of certainty to be aimed at must depend on the pecuniary advantage to be gained by following the result of the experiment, compared with the increased cost of the new method, if any, and the cost of each experiment (Gosset 1905a; reprinted in Pearson (1939)).[34]

For example, setting the optimal level of significance is not to be done conventionally or by "some outside authority in mathematics", Student said from the beginning of his statistical inquiries.

[Figure 1 about here]

[34] Compare Savage (1954, p. 116): "In principle, if a number of experiments are available to a person, he has but to choose one whose set of derived acts has the greatest value to him, due account being taken of the cost of observation". Contrast the purpose of experiments according to Gosset, Pearson, Jeffreys, and Savage with that of Fisher (1926, p. 504): "Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level" (quoted in Ziliak and McCloskey 2008, p 46).

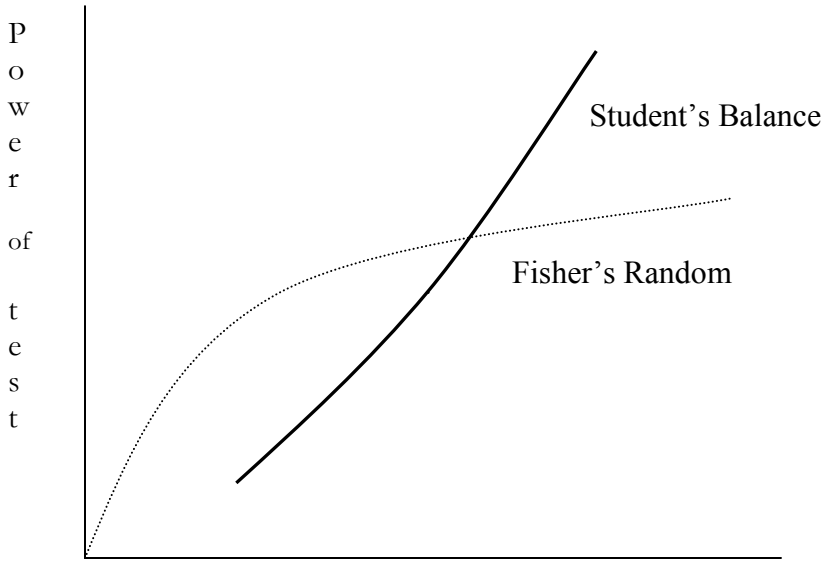

Figure 1. Size of real treatment difference measured by the variance of Student's Δ's

7. The third wave thesis

Student's simple yet exceedingly important point is that fertility contours in the land are non-random, they are systematic and regular, and this fact requires a degree of freedom to capture non-random plot-and-block specific contributions to error terms of mean squares of ANOVAs (Student 1923, p. 103). The general point that Student proved and Fisher chose to ignore during the "first wave" of field experiments is that artificial randomization gives less assurance, not more, when observations are spatially and/or temporally correlated and confounded by one or more variables.

An important result follows: if real treatment variance is large with respect to the total error of the experiment, balanced designs will detect the large effects with higher probability than random. But when real effects from treatments are small – when the experimental treatment is not much different from the company or industry standard (the control) - random designs generate "significant" effects more often. But as Student showed, the latter result is spurious, comparatively unpromising from the economic viewpoint, and higher in both total and marginal costs relative to balanced alternatives.

Following Fisher's model, the Levitt and List study asserts in a single sentence only that "randomization bias is not a major empirical problem for field experiments of the kind we conducted" (L and L, p. 14) offering no evidence or explanation for their belief. Moreover, Levitt and List do not discuss the vast body of evidence accumulated against completely randomized blocks during the "first wave" of field experiments, beginning with Student's evidence.

Levitt and List are hardly alone. Varian (2011), the Chief Economist of Google, claims that randomized trials are "the gold standard for causal inference" and "ran about 6,000" of them

in 2010—perhaps some of them on economists, unaware. Duflo claims that field experimentalists have borrowed from medicine a "'very robust and very simple tool' . . . they subject social policy ideas to randomized control trials, as one would use in testing a drug. 'This approach,' Duflo claims, 'filters out statistical noise; it connects cause and effect'" (quoted in Parker 2010, pp. 79-80).

But that is not generally the case; that assertion has not been proven by the new field experimentalists. Randomization plus a low level of Type I error is neither necessary nor sufficient for proving causality or advancing the economics of the firm. As Heckman and Vytlacil (2007) put it, "Randomization is . . . not an ideal or 'gold standard'". Fisher, an academic statistician, was well aware of this; but from 1925 and ever after, as Ziliak and McCloskey have shown, Fisher refused to acknowledge Student's economic and indeed beeronomic motives for experimenting in the first place.

Yet in the private sector of the economy – where Student worked – experimentation is not an academic problem; it is a business problem. Abstract validity and spurious "significance" give way to profit and loss, quality assurance, and other substantive business goals. In the early 1900s, for example, the Guinness Brewery purchased more than 1 million barrels of malted barley annually, consuming in the form of Guinness stout about one-third of the annual Irish barley crop.[35] Each time a new variety of barley or malt was introduced in bulk, the marginal cost to the firm could rise to millions of pounds sterling—and with the added uncertainty of brewing suboptimal beer. Student needed confidence that observed treatment differences were real and profitable. He needed to know the odds that an observed difference, however large, represented a real and meaningful difference. As Beaven (1947, p. 293) observed long ago, "[m]any of the 'randomized' plot arrangements appear to be designed to illustrate statistical theory . . . only a trifling number of them so far have demonstrated any fact of value to the farmer".

Balanced designs are thus more, not less, valid in the old Latin sense of *validus* and *valere*, meaning "robustness", "strength" and "value" (Shorter Oxford English Dictionary, 2002).

Evidently the authors' third and final claim—the third wave thesis—must also be reconsidered. As a matter of chronological fact, Student combined laboratory and field experiments in an industrial organization a full century before Levitt's and List's "third wave". The theory of the firm is not likely to advance much by the addition of more randomized trials; balanced and repeated trials are, by contrast, more profitable and precise.[36]

---

[35] "Comparative Statement of Financial Operations," Arthur Guinness Son & Co. Ltd., GDB C004.06/0016), Guinness Archives, Diageo (Dublin); the figure does not include imports of barley and malt from abroad.

[36] An anonymous referee questioned two of the more general claims made by Levitt, List, and others such as Banerjee and Duflo (2011): (1) that experiments are different from, and superior to, observations; and (2) that new field experiments are meaningfully different from field and laboratory experiments of the past. The referee called both claims "artificial"—a view that might be shared by Student (1908, pp. 1-2) and Savage (1954, p. 118), for example. Said Savage (1954, p. 118): "Finally, experiments as opposed to observations are commonly supposed to be characterized by reproducibility and repeatability. But the observation of the angle between two stars is easily repeatable and with highly reproducible results in double contrast to an experiment to determine the effect of exploding an atomic bomb near a battleship. All in all, however useful

That is the scientific finding again and again, in economics as much as in epidemiology and medicine, as Heckman and Vytlacil and Deaton and Altman and Rothman and many others have shown. Balanced designs have shown their higher power in discrete choice models by Carson, Louviere, and Wasi (2009), in agricultural economics, and by Bruhn and McKenzie in their massive survey comparing balanced and randomized designs in development projects related to the World Bank.

Between the poles of perfect balance and pure randomness there exists of course a world of possible compromise. Fisher's Latin square is, for example, both random and balanced "thus conforming to all the principles of allowed witchcraft" (Student 1938, p. 365).

Field experiments will continue to thrive, "fooled by randomness," as Taleb (2005) puts it. Still it might pay to reflect on an article by Stigler (1982 [1969], p. 107), who foresaw from the annals of economic thought the current illusion of history and theory:

> The young theorist, working with an increasingly formal, abstract, and systematic corpus of knowledge, will . . . assume that all that is useful and valid in earlier work is present—in purer and more elegant form—in the modern theory.

---

the distinction between observation and experiment may be in ordinary practice, I do not yet see that it admits of any solid analysis".

References

Altman, D. G., Schultz, K.F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gotzsche, P., Lang, T. 2001. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Annals of Internal Medicine 134, 663-691.

Banerjee, A., Duflo, E. 2011. Poor economics: a radical rethinking of the way to fight global poverty. Public Affairs, New York.

Beaven, E. S. 1947. Barley: fifty years of observation and experiment. Duckworth, London.

Box, G. E.P., Hunter, W.G., Hunter, J.S. 1978 [2005]. Statistics for experimenters. John Wiley & Sons, New York.

Bruhn, M., McKenzie, D. 2009. In pursuit of balance: randomization in practice in development economics. American Economic Journal: Applied Economics 1,: 200–232.

Carson, R.T., Louviere, J.J., Wasi, N. 2009. A cautionary note on designing discrete choice experiments. American Journal of Agricultural Economics 91, 1056–1063.

Cochrane, W.G. 1976. Early development of techniques in comparative experimentation. In: Owen, D.B. (Ed.), On the History of Statistics and Probability. Marcel Dekker Inc., New York, p. 126.

Cochrane, W. G. 1989. Fisher and the analysis of variance. Pp. 17-34 in Fienberg S.E., Hinkley, D.V., eds., R. A. Fisher: an appreciation. Springer-Verlag, New York.

Concise Dictionary of National Biography, Part II, 1901-1950. Oxford, Oxford University Press.

Deaton, A. 2007. Evidence-based aid must not become the latest in a long string of development fads. In: Banerjee, A. (Ed.), Making aid work. MIT Press, Cambridge, pp. 60-61.

Duflo, E., Glennerster, R., Kremer, M. 2007. Using randomization in development economics research: a toolkit, MIT Department of Ecconomics and J-PAL Poverty Action Lab.

Es, H.M., Gomes, C.P., Sellman, M., van Es, C.L. 2007. Spatially-balanced complete block designs for field experiments. Geoderma 2007 140, 346-352.

Fisher, R.A., 1922. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London A 222, 309–368

Fisher, R. A., Mackenzie, W.A. 1923. Studies in crop variation: II. The manurial response of different potato varieties. Journal of Agricultural Science 13, 311–320.

Fisher, R. A. 1925 [1928]. Statistical methods for research workers. G.E. Stechart, New York.

Fisher, R. A. 1926. Arrangement of field experiments. Journal of Ministry of Agriculture 33, 503-13.

Fisher, R.A. 1933. The contributions of rothamsted to the development of the science of statistics. In: Rothamsted Experimental Station, Annual report. Rothamsted, Rothamsted, pp. 43-50.

Fisher, R. A. 1935. The design of experiments. Oliver & Boyd, Edinburgh.

Fisher, R. A. 1955. In: Mendel, G. (Ed.) Experiments in plant hybridisation. Oliver & Boyd, Edinburgh, p. 6.

Friedman, M. 1953. The effects of a full-employment policy on economic stability: a formal analysis. In: Friedman, M., Essays in positive economics. University of Chicago Press, Chicago, pp. 117-132.

Gosset, W. S. aka Student. [See below: Student]. 1904. The application of the 'law of error' to the work of the brewery. Guinness Laboratory Report 8. Arthur Guinness & Son, Ltd., Guinness Archives, Dublin, pp. 3–16 and unnumbered appendix.

Gosset, W. S. 1905a. Letter to Karl Pearson, Guinness Archives (Dublin), GDB/BRO/1102 (partially reprinted in Pearson 1939, pp. 215-216).

Gosset, W. S. 1905b. The Pearson co-efficient of correlation. Guinness Laboratory Report 3. Arthur Guinness & Son, Ltd., Guinness Archives, Dublin.
Gosset, W.S. 1936. Co-operation in large-scale experiments. Supplement to the Journal of the Royal Statistical Society 3, 115-36.
Gosset, W. S. 1962. Letters of William Sealy Gosset to R. A. Fisher. Vols. 1-5, Eckhart Library, University of Chicago. Private circulation.
Hall, A. D. 1905. The book of rothamsted experiments. E.P. Dutton and Company, New York.
Harrison, G. W., List, J.A. 2004. Field experiments. Journal of Economic Literature 44, 1009-1055.
Heckman, J.J. 1991. Randomization and social policy evaluation. NBER Working Paper No. T0107. National Bureau of Economic Research, Cambridge.
Heckman, J.J., Smith, J. 1995. Assessing the case for social experiments. Journal of Economic Perspectives 9, 85-110.
Heckman, J.J., Vytlacil, E.J. 2007. Econometric evaluation of social programs, part I: causal models, structural models and econometric policy evaluation. In: Heckman, J.J., Leamer, E. (Eds.) Handbook of econometrics 6B. Elsevier, Amsterdam, pp. 4836.
Herberich, D. H., S. D. Levitt, List, J.A. 2009. Can field experiments return agricultural economics to the glory days? American Journal of Agricultural Economics 91, 1259-1265.
Horace. 20 B.C. Epistle I.19, to Maecenas. In: Satire and Epistles. Smith Palmer Bovie (Transl.). University of Chicago Press, Chicago, p. 220.
Jeffreys, H. 1939 [1961]. Theory of probability. Oxford University Press, London. Third revised edition.
J-PAL Poverty Action Lab. 2010. When did randomized evaluations begin? Poverty Action Lab, MIT, Cambridge. http://www.povertyactionlab.org/methodology/when/when-did-randomized-evaluations-begin
Karlan, D., List, J. 2007. Does price matter in charitable giving? Evidence from a large-scale natural field experiment. American Economic Review 97, 1774-1793.
Karlan, D., Appel, J. 2011. More than good intentions: How a new economics is helping to solve global poverty. Dutton, New York.
Kruskal, W. H. 1980. The significance of fisher: a review of R. A. Fisher: The life of a scientist. Journal of the American Statistical Association 75, 1019-30.
Lanham, R.A. 1991. A handlist of rhetorical terms. University of California Press, Los Angeles.
Leon, A.C., Demirtas, H., Hedeker, D. 2007. Bias reduction with an adjustment for participants' intent to drop out of a randomized controlled clinical trial. Clinical Trials 4, 540-547.
Levitt, S. D., List, J.A. 2009. Field experiments in economics: the past, the present, and the future. European Economic Review 53, 1-18.
List, J. A. 2009. An introduction to field experiments in economics. Journal of Economic Behavior and Organization 70, 439-442.
List, J.A., Schogren, J. 1998. Calibration of the difference between actual and hypothetical valuations in a field experiment. Journal of Economic Behavior and Organization 37, 193-205.
McCloskey, D. N., Ziliak, S.T. 2009. The unreasonable ineffectiveness of fisherian 'tests' in biology, and especially in medicine. Biological Theory 4, 44-53.
Mercer, W. B., Hall, A.D. 1911. The experimental error of yield trials. Journal of Agricultural Science 4, 107-127.

Moore, D., McCabe, G. 1998. Introduction to the practice of statistics. W. H. Freeman, New York. Third edition.
Neyman, J. 1938. Mr. W. S. Gosset. Journal of the American Statistical Association 33, 226-228.
Neyman, J. 1961. Silver jubilee of my dispute with Fisher. Journal of Operations Research 3, 145-54.
Neyman, J., Iwaszkiewicz, K., Kolodziejczyk, S. 1935. Statistical problems in agricultural experimentation. Supplement to the Journal of the Royal Statistical Society 2, 107-80.
Neyman, J., Pearson, E.S. 1938. Note on some points on 'Student's' paper on 'comparison between balanced and random arrangements of field plots'. Biometrika 29, 379-88.
Parker, I. 2010. The Poverty Lab. The New Yorker, May 17, 79-80.
Pearson, E. S. 1938. Some aspects of the problem of randomization: II. An illustration of Student's inquiry into the effect of 'balancing' in agricultural experiment. Biometrika 30, 159-179.
Pearson, E. S. 1939. 'Student' as statistician. Biometrika 30, 210-50.
Pearson, E. S. 1990 [posthumous]. 'Student': a statistical biography of William Sealy Gosset. Clarendon Press, Oxford. Edited and augmented by R. L. Plackett, with the assistance of G. A. Barnard.
Peirce, C.S., Jastrow, J. 1885. On small differences of sensation. Memoirs of the National Academy of Sciences for 1884 3, 75-83.
Press, S. J. 2003. Subjective and objective bayesian statistics. Wiley, New York.
Reid, C. 1982 [1998]. Neyman: a life. Springer, New York.
Rodrik, D. 2008. The new development economics: we shall experiment, but how shall we learn? Brookings Development Conference, Harvard University, John F. Kennedy School of Government.
Rothman, K., Greenland, S., Lash, T. 2008. Modern epidemiology. Lippincott, Williams & Wilkins, Philadelphia.
Rubin, D. 1990. Comment: Neyman (1923) and causal inference in experiments and observational studies. Statistical Science 5, 472–480.
Savage, L. J. 1954. The foundations of statistics. Dover, New York.
Savage, L. J. 1971 [1976 posthumous]. On re-reading R. A. Fisher. Annals of Statistics 4, 441-500.
Shorter Oxford English Dictionary. 2002. Valid, validity. Oxford University Press, Oxford, p. 3499.
Splawa-Neyman, J., 1923 [1990]. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Statistical Science 5, 465–472. Translated from Polish to English in 1990, by Dabrowska, D.M, Speed, T.P. (Eds.).
Stigler, G. J. 1969 [1982]. Does economics have a useful past? History of Political Economy 2, 217-230. In: Stigler, G. J., The economist as preacher. University of Chicago Press, Chicago, pp. 107-118.
Stigler, S. M. 1986. The history of statistics: the measurement of uncertainty before 1900. Harvard University Press, Cambridge.
Street, D. 1990. Fisher's contributions to agricultural statistics. Biometrics 46,937–945.
Student. 1907. On the error of counting with a haemacytometer. Biometrika 5, 351-60.
Student. 1908. The probable error of a mean. Biometrika 6, 1-24.

Student. 1911. Appendix to Mercer and Hall's paper on 'The experimental error of field trials'. Journal of Agricultural Science 4, 128-131. In: Student. Student's collected papers, pp. 49-52. Pearson, E. and J. Wishart, (Eds.)
Student. 1923. On testing varieties of cereals. Biometrika 15, 271-293.
Student. 1925. New tables for testing the significance of observations. Metron 5, 105-108.
Student. 1926. Mathematics and agronomy. Journal of the American Society of Agronomy 18. Reprinted in: E. S. Pearson, Wishart, J. (Eds.), Student's Collected Papers (1942). University College London, London, pp. 121-34.
Student. 1927. Errors of routine analysis. *Biometrika* 19, 151-64
Student. 1931a. The Lanarkshire milk experiment. *Biometrika* 23, 398-406.
Student. 1931b. Yield trials. In: Bailliere's encyclopedia of scientific agriculture, pp. 1342-1360. Reprinted in: Pearson, E.S., Wishart, J. (Eds.) Student's collected papers (1942), University College London, London, pp. 150-168.
Student. 1936. The half-drill strip system. Letter to Nature 138, 971.
Student. 1938 [posthumous]. Comparison between balanced and random arrangements of field plots. Biometrika 29, 363-78.
Student. 1942 [posthumous]. Student's collected papers. University College London, London. E. S. Pearson, Wishart, J. (Eds.).
Taleb, N.N. 2005. Fooled by randomness: the hidden role of chance in life and in the markets. Random House, New York.
Varian, Hal. 2011. Are randomized trials the future of economics? Federalism offers opportunities for casual [sic] experimentation. The Economist, April 27th. http://www.economist.com/node/21256696
Wood, T. B., Stratton, F. J.M. 1910. The interpretation of experimental results. Journal of Agricultural Science 3, 417-440.
Wright, R. 1998. Haiku: this other world. Random House, New York.
Yates, F. 1964. Sir Ronald Fisher and the design of experiments. Biometrics 20, 307-321.
Zellner, A. 2004. Statistics, econometrics, and forecasting. Cambridge University Press, Cambridge.
Zellner, A., Rossi, P. 1986. Evaluating the methodology of social experiments. Proceedings of the Federal Reserve Bank of Boston, 131-166.
Ziliak, S. T. 2008. Guinnessometrics: the economic foundation of 'Student's' *t*. Journal of Economic Perspectives 22, 199-216.
Ziliak, S. T. 2010. The *validus medicus* and a new gold standard. The Lancet 376, 324-325.
Ziliak, S.T. 2011a. Statistical significance on trial: a significant rejection by the supreme court. Significance 8, forthcoming, September issue. Royal Statistical Society and American Statistical Association, London and Washington, DC.
Ziliak, S. T. 2011b. W.S. Gosset and some neglected concepts in experimental statistics: Guinnessometrics II. Journal of Wine Economics, forthcoming.
Ziliak, S. T., McCloskey, D.N. 2008. The cult of statistical significance: how the standard error costs us jobs, justice, and lives. University of Michigan Press, Ann Arbor.

Research Papers
2011

2011-11: Eduardo Rossi and Paolo Santucci de Magistris: Estimation of long memory in integrated variance

2011-12: Matias D. Cattaneo, Richard K. Crump and Michael Jansson: Generalized Jackknife Estimators of Weighted Average Derivatives

2011-13: Dennis Kristensen: Nonparametric Detection and Estimation of Structural Change

2011-14: Stefano Grassi and Paolo Santucci de Magistris: When Long Memory Meets the Kalman Filter: A Comparative Study

2011-15: Antonio E. Noriega and Daniel Ventosa-Santaularia: A Simple Test for Spurious Regressions

2011-16: Stefano Grassi and Tommaso Proietti: Characterizing economic trends by Bayesian stochastic model specification search

2011-17: Søren Johansen and Theis Lange: Some econometric results for the Blanchard-Watson bubble model

2011-18: Tom Engsted and Thomas Q. Pedersen: Bias-correction in vector autoregressive models: A simulation study

2011-19: Kim Christensen, Roel Oomen and Mark Podolskij: Fact or friction: Jumps at ultra high frequency

2011-20: Charlotte Christiansen: Predicting Severe Simultaneous Recessions Using Yield Spreads as Leading Indicators

2011-21: Bent Jesper Christensen, Olaf Posch and Michel van der Wel: Estimating Dynamic Equilibrium Models using Macro and Financial Data

2011-22: Antonis Papapantoleon, John Schoenmakers and David Skovmand: Efficient and accurate log-Lévi approximations to Lévi driven LIBOR models

2011-23: Torben G. Andersen, Dobrislav Dobrev and Ernst Schaumburg: A Functional Filtering and Neighborhood Truncation Approach to Integrated Quarticity Estimation

2011-24: Cristina Amado and Timo Teräsvirta: Conditional Correlation Models of Autoregressive Conditional Heteroskedasticity with Nonstationary GARCH Equations

2011-25: Stephen T. Ziliak: Field Experiments in Economics: Comment on an article by Levitt and List