Review of Behavioral Economics, 2014, 1: 167–208

Balanced versus Randomized Field Experiments in Economics: Why W. S. Gosset aka "Student" Matters

Stephen T. Ziliak*

Roosevelt University, 430 S Michigan Ave Chicago, IL 60605, United States; sziliak@roosevelt.edu

ABSTRACT

Over the past decade randomized field experiments have gained prominence in the toolkit of empirical economics and policy making. In an article titled "Field Experiments in Economics: The Past, the Present, and the Future," Levitt and List (2009) make three main claims about the history, philosophy, and future of field experiments in economics. (1) They claim that field experiments in economics began in the 1920s and 1930s in agricultural work by Neyman and Fisher. (2) They claim that artificial randomization is essential for good experimental design because, they claim, randomization is the only valid justification for Student's test of significance. (3) They claim that decision-making in private sector firms will be advanced by partnering with economists doing randomized experiments. Several areas of research have been influenced by the article despite the absence of historical and methodological

* I thank the editor, three referees, and numerous colleagues for helpful comments on a previous version of this manuscript. Comments from Doug Altman, Gary Charness, Angus Deaton, Harold van Es, Nathan Fiala, Steven Goodman, Glenn Harrison, James Heckman, Tim Harford, Dani Rodrik, Mark Thoma, David Wofford, Allan Wurtz, the late Arnold Zellner, and James P. Ziliak were especially helpful. For permission to publish from archival material, I thank the principals at the Guinness Archives (Diageo) and at the Special Collections Library of University College London.

ISSN 2326-6198; DOI 10.1561/105.00000008

review. This paper seeks to fill that gap in the literature. The power and efficiency of balanced over random designs — discovered by William S. Gosset aka Student, and confirmed by Pearson, Neyman, Jeffreys, and others adopting a balanced, decision-theoretic and/or Bayesian approach to experiments — is not mentioned in the Levitt and List article. Neglect of Student is regrettable. A body of evidence descending from Student (1911) and extending to Heckman and Vytlacil (2007) suggests that artificial randomization is neither necessary nor sufficient for improving efficiency, identifying causal relationships, and discovering economically significant differences. Three easy ways to improve field experiments are proposed and briefly illustrated.

Keywords: Field experiments; statistical significance, Levitt, List.

JEL Codes: C93, C9, B1

> 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

Over the past decade randomized field experiments have gained prominence in the toolkit of empirical economics and policy making. In an article titled "Field Experiments in Economics: The Past, the Present, and the Future," Levitt and List (2009) make three main claims about the history, philosophy, and future of field experiments in economics.

Claim 1 says that field experiments in economics began "in the 1920s and 1930s" in agricultural research by the mathematical statisticians Jerzy Neyman and Ronald A. Fisher.[1]

Claim 2 says that the introduction of randomization — of completely randomized blocks — laid the "foundation" for good experimental design in both economics

[1] Levitt and List (2009, pp. 1, 3–5).

and statistics.[2] Artificial randomization of treatments and controls is said by the authors to provide the only "valid" justification for use of Student's test of statistical significance.[3]

An example of artificial randomization is the use of computer, dice, shuffled cards, or other random number generator to allocate treatments and controls to experimental units. For example, a potato farmer 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 randomization the authors credit "Splawa-Neyman (1923 [1990])" and "Fisher and McKenzie (1923)" [sic] — these articles, Levitt and List claim, established the "experimental foundation" which the authors equate with randomization.[4]

Claim 3 about the past, present, and future of field experiments in economics states that the current generation of field experimentalists has seen "much deeper" than "previous generations".[5] Thus the theory of the firm will advance just as, the authors assume, the fields of medicine and pharmacology have so advanced, with economists conducting randomized trials for private sector firms.[6] 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".[7] "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".[8] "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 also permit a

---

[2] Levitt and List (2009, p. 3).

[3] Levitt and List, 2009, p. 3; contrast Ziliak (2008, 2010, 2011b), Ziliak and McCloskey (2008), and the unanimous rejection of statistical significance by the Supreme Court of the United States in Matrixx Initiatives, Inc. v. Siracusano, No. 09-1156, on March 22, 2011 (Ziliak, 2011a).

[4] Levitt and List, 2009, pp. 1, 2–4; Splawa-Neyman and Neyman are the same person — Jerzy Neyman. He shortened his name in the 1920s, when he moved from Poland to England.

[5] Levitt and List, 2009, p. 15.

[6] But see the revised CONSORT statement on randomized controlled trials in medicine by Altman *et al.* (2001) and also Rothman *et al.* (2008); for discussion of Student's pioneering comparisons of random and balanced experiments see Ziliak (2011a, 2011b, 2010, 2008) and Ziliak and McCloskey (2008, Ch. 1, 20–23). See also: Harrison (2011), Rodrik (2008).

[7] Levitt and List, 2009, p. 2; see Ziliak (2008), which shows that private sector use of randomized field experiments in economics began with Gosset's work for the Guinness Brewery more than a century ago.

[8] Levitt and List, 2009, p. 18.

deeper empirical exploration into problems that excite economists, practitioners, and policymakers".[9]

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.

2 Recent Literature

To date no study has endeavored to assess their history and philosophy of randomization and field experiments in economics. This is unfortunate but, on second thought, not surprising. As Deaton (2007) 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.[10] "I am a huge fan of randomized trials," Varian (2011) told *The Economist,* "they are well worth doing since they are the gold standard for causal inference".

Others, citing Fisher (1925, 1935), appear to agree. Yet the new generation of field experimentalists — notably, Levitt and List (2009), Karlan and Appel (2011) and Duflo, Banerjee, and others at the Abdul Latif Jameel Poverty Action Lab (J-PAL, 2010) — join Varian and assume rather than prove the validity of complete randomization. The alleged comparative advantage of randomized trials has not been established by comparative testing with balanced and other designs — despite an enormous body of published literature comparing balanced and random designs from Student (1911) to Heckman and Vytlacil (2007).

The authority of today's randomization school seems to derive from uncritical acceptance of assertions by Ronald A. Fisher in *The Design of Experiments* (1935) and *Statistical Methods for Research Workers* (1925). "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."[11] Levitt and List consider this quote from Fisher (1935) the foundation of experimental method:

---

[9] Levitt and List, 2009, p. 2; thus List and Shogren (1998) is an early contribution to the third wave, so conceived.

[10] Deaton (2007, p. 25); Levitt and List (2009, pp. 2, 15–18). Google Scholar credits Levitt and List (2009) with 124 citations as of May 29, 2013.

[11] Levitt and List (2009, p. 3).

> 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 [via complete randomization of treatments, controls, and varieties] 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.[12]

To Fisher units have the same probability of being "treated alike" only when treatments and controls are randomly assigned. Yet Levitt and List — following Cochrane (1976), Rubin (1990), and Street (1990) — assume rather than prove that Fisher "laid the experimental foundation" with his randomization thesis.

Most economists do not know how to judge the three main claims made by Levitt and List (2009). Most are not equipped to judge their sweeping assertions about randomization and the history of field experiments in economics and statistics.[13]

This paper draws on a century of theory and evidence not examined by the new field experimentalists: Gosset (1904, 1905a, 1905b, 1936), Student (1908, 1911, 1923, 1938, 1942), Wood and Stratton (1910), Mercer and Hall (1911), Pearson (1938, 1990), Neyman (1934, 1938), Neyman *et al.* (1935), Neyman and Pearson (1938), Beaven (1947), Savage (1954, 1976), Jeffreys (1939 [1961]), Kruskal (1980), Zellner and Rossi (1986), Heckman (1991), Berk and Freedman (2003), Heckman and Vytlacil (2007), Ziliak (2008), Ziliak and McCloskey (2008), Carson *et al.* (2009), Bruhn and McKenzie (2009), and others preceding Levitt and List (2009).[14]

It turns out that in the 1910s, 1920s, and 1930s — the authors did not mention — Student (1911, 1923, 1938), Mercer and Hall (1911), Gosset (1936), Pearson (1938, 1990), Neyman *et al.* (1935), Neyman (1938), Jeffreys (1939, [1961]) and others proved the opposite claim — they proved that balanced designs are more precise and more economically efficient than randomization — particularly when

---

[12] Fisher (1935, p. 26), quoted by Levitt and List (2009, p. 3).

[13] Herberich *et al.* (2009) make similar erroneous assertions about the history and development of field experiments in agricultural economics.

[14] Though Levitt and List mention Student (1923) and Heckman and Vytlacil (2007) their article does not mention that these and the other studies are evidence against randomization and in favor of balance. For discussion of Fisher's methods, and how they differ from the methods of Student, Egon Pearson, Harold Jeffreys, and other experimental economists — the decision-theorists and Bayesian experimentalists not mentioned by Levitt and List — see Zellner (2004), Ziliak (2008, 2010, 2011b), Ziliak and McCloskey (2008, Chapters 1, 17–23), McCloskey and Ziliak (2009), and Harrison (2011).

economic stakes and confounding are large, overturning the so-called principle of randomization.[15]

From Student (1911, 1923) and Beaven (1947) to Jeffreys (1939 [1961]) and Heckman (1991), leading experimentalists and economic statisticians have taken a balanced and economic approach to the logic of experimental uncertainty, demonstrating with balanced designs both higher precision and lower cost, thereby refuting Fisher (1935, 1925) and the completely randomized blocks that Levitt, List, and others currently endorse.

Neglect of Student is unfortunate. The misfortune is not that Student has priority over Neyman and Fisher, though, as I have shown, he does. From 1905 to 1937, William S. Gosset aka Student (1876–1937) showed by repeating field experiments on barley and other crops in a profit-seeking environment — in his job as Head Experimental Brewer and finally as Head Brewer of Guinness — that artificial randomization of treatments and controls is neither necessary nor sufficient for improving efficiency, identifying causal relationships, and discovering economically significant differences.[16]

Student's success with balanced designs delighted Neyman (1938), Neyman *et al.* (1935), Pearson (1990, 1938), and Jeffreys (1939), and troubled Fisher (1935, 1925). Thus the endorsement of Fisher to the neglect of Student comes with a cost.

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.[17] Balanced designs are deliberate or systematic arrangements of treatments and controls when, observed or not, there is known to be one or more non-random sources of fluctuation confounding the output of interest (the dependent variable). For example, when testing the yield of barley A against barley B, balancing is necessary when a differential fertility gradient cuts systematically across the farm field, affecting the size and variance of the experimental unit, as it

---

[15] Student (1938) and Pearson (1938) simulate alternative models confirming the economic advantage of balance, a fact Neyman (1935, 1938) believed was established rigorously enough by Student (1923). See also: Jeffreys (1939 [1961], pp. 243–244).

[16] Gosset was a self-trained innovator of experimental design and analysis, skills he learned on the job at Guinness's brewery, where he worked for 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, 2011a, 2011b, 2008, for discussion).

[17] Student's legacy at the frontier is extensive, though not always acknowledged: Pearson (1938, 1939, 1990), Neyman *et al.* (1935), Neyman (1938), Jeffreys (1939 [1961]), Savage (1954, 1976), Deming (1978), Kruskal (1980), Zellner and Rossi (1986), Heckman (1991), Press (2003), and Heckman and Vytlacil (2007), to name a few.

always does (Student, 1911, 1923, 1938; van Es *et al.*, 2007). Balancing is necessary when unobserved attitudes about schooling, gender, and race contribute to selection bias in the 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 fluctuation, chance and systematic, random and real; thus balancing is more, not less, valid than complete randomization (Ziliak, 2011b).

As Student observed, in short, the need to balance is present any time there is a secular or temporal correlation of the dependent variable (such as crop yield) with one or more systematic, non-random independent variables (such as soil fertility) which cannot be artificially randomized.

The article does not do justice to Student's pioneering work and it fails to discuss the most important controversy in the twentieth century history of experimental design — the Student–Fisher debates of the 1920s and 1930s.[18] There are numerous other, more minor errors in the Levitt–List history and methodology that will not be discussed in this article.[19]

3 Note on Methods and Sources

The method here is textual exegesis of primary and secondary literature mentioned and not mentioned in the history by Levitt and List (2009). The analysis of published material is supplemented by evidence found in unpublished Gosset archives owned and copyrighted by Guinness Archives (Diageo), Dublin, Ireland, and by the Special Collections Library, University College London, UK. Some of the discussion of Student is based on unpublished brewing material found in Guinness *Laboratory Reports*, 1898–1912, but nothing asserted here about the Levitt and List (2009) history and methodology depends on unpublished material. Still, the laboratory reports contain some of Gosset's (Student's) most important theory and results on small samples of repeated experiments in the laboratory and field (see Gosset, 1904). Other primary sources include over 150 letters from Gosset to Fisher (Gosset, 1962)

---

[18] See Student (1938), for example, and Ziliak and McCloskey (2008, Chapters 20–22).

[19] Levitt and List (2009, p. 4), for example: Gosset (spelled with one t as in one t-distribution) was not studying yeast when he discovered the need for a small sample distribution (Ziliak, 2008). And it is not true that Guinness employees could not publish research. The proprietary conditions were that they publish under a pen name and that they avoid discussing beer and Guinness. In *Biometrika* alone, between 1907 and 1937, there were numerous articles published by Guinness scientists: 14 articles by Student alone, and one or more each by his assistants, "Mathetes," "Sophister," and others (Student, 1942; Ziliak and McCloskey, 2008, pp. 213, 217).

and: unpublished scientific notebooks, memoranda, other correspondence, annual reports, and financial statements.[20]

4 The First Wave Thesis

The first mistake is chronological. The authors claim that field experiments in economics began with Neyman and Fisher in the 1920s and 1930s:

> 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.[21]

But this claim is easily dismissed. If the goal is to identify first examples of field experiments in agricultural economics the clock of history should start much earlier than Levitt and List claim. Consider, for example, *The Farmer's Letters to the People of England* (1767), by Arthur Young — published 250 years ago. More importantly, given the authors' statistical focus, there was more than a decade before Fisher (1925) and Splawa-Neyman (1923) pioneering work done by Wood and Stratton (1910), Mercer and Hall (1911), and especially Student (1911, 1923).

Neglect of Student is particularly unfortunate as inclusion of Student reveals weighty evidence against each of the three main claims asserted by Levitt, List, and by now many others in experimental economics.

After Student's (1908) introduction of the small sample table and test of significance to modern statistics, a better start for the "first wave" of field experiments in economics is Student's (1911) appendix to the Mercer and Hall (1911) experiment. Student's pioneering work with Mercer and Hall was published 79 years before Splawa-Neyman (1923) was first translated into English, and Student had a far greater impact on the origin and development of modern experiments.[22]

---

[20] Gosset was a prolific letter writer and many leading scientists — from Karl Pearson to Jerzy Neyman — sought his counsel. For example, Gosset (1962) is a five-volume book containing more than 150 letters from Gosset to Fisher, written and exchanged between 1915 and 1934.

[21] Levitt and List, 2009, p. 1.

[22] Student (1911) is reprinted in *Student's Collected Papers* (1942) but neither is Student (1942) mentioned by Levitt and List.

Mercer and Hall (1911, p. 127) thank Student for designing and evaluating the results of a field experiment on mangolds and wheat which they 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.

The "Hall" of Mercer and Hall is A. Daniel Hall, a noted experimentalist whose main job was Director of the famous Rothamsted Agricultural Experimental Station, where Fisher would later work (Hall, 1905). In British scientific circles, Student was by 1911 well-known for his *Biometrika* article on small sample theory (Student, 1908). Student's unpublished work with the experimental maltster and barley farmer Beaven (1947), on the design and evaluation of field experiments, was also familiar to Hall, a long-time associate of Beaven and others affiliated with the Guinness brewery.

Thus the Guinness brewer, Student, was recruited by Hall for assistance. Student titled his 1911 article "Note on a Method of Arranging Plots so as to Utilize a Given Area of Land to the Best Advantage in Testing Two Varieties".[23] Using Mercer's and Hall's mangolds data, Student found that standard deviations of mean yield differences are reduced as the experimental block size is reduced. The reason, Student found, is a non-random, systematic, and confounding variable, spoiling artificially randomized arrangements: that is, Student argued from 1911, the non-random occurrence of fertility gradients in the soil.[24] The correlation of neighboring blocks, caused by the changing marginal productivity of the soil as one travels from one block to another (and even from inch to inch) must figure into the design of the experiment, Student discovered, or else an important non-random source of crop growth will bias results.

Student made several other advances in the 1911 article. More than two decades before Fisher (1935) wrote about "the concepts of repetition, blocking, and randomization" (Levitt and List, 2009, p. 3) Student found through repetition, blocking, and in comparisons of random with balanced designs that the closer

[23] See also: Student (1923); Gosset (1936); and Beaven (1947, pp. 238, 254–255).
[24] See also: Student (1938); compare Pearson (1938, 1990).

in space competing crop varieties and/or treatments are sown (through blocking or pairing — what Student and Beaven called "the principle of maximum contiguity") the more "real" are the observed differences between varieties (or between treatments and controls, if that is what is being tested).[25]

4.1 Field Experiments in Economics Before Fisher, 1911 to 1923

Student (1911) used blocking and stratification before the words existed in a statistical sense. As Deming (1978, p. 879) noted, "Stratification is equivalent to blocking in the design of an experiment."

What is a block? Box *et al.* (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." Blocks are strata. Thus it is fair to say that Student (1911, 1938) used and advocated blocking and stratification from 1911 forward, through three decades of pioneering work on field experiments in economics not mentioned by Levitt and List.

Deming (1978), who did a long stint at the U.S. Department of Agriculture prior to his career in manufacturing, agreed with Student: complete random sampling and randomized experiments are at best preliminary steps to scientific study. Complete randomization has a purpose when the investigator knows little or nothing at all about strata and when the cost of being wrong is negligible. Said Deming (1978, p. 879):

> The primary aim of stratified sampling is to increase the amount of information per unit of cost. A further aim may be to obtain adequate information about certain strata of special interest.
>
> One way to carry out stratification is to rearrange the sampling units in the frame so as to separate them into classes, or strata, and then to draw sampling units from each class. The goal should be to make each stratum as homogeneous as possible, within limitations of time and cost.[26]

---

[25] Student (1911 [1942], p. 52); Beaven (1947, pp. 264, 273–275).

[26] Deming (1978, p. 879). Deming said he learned it from Neyman (1934). In the seminal article Neyman proves mathematically and empirically the statistical and economic advantages of

And that is exactly what Student (1911) showed. In his notable book, *Planning of Experiments*, Cox (1958) recommends "completely randomized arrangement . . . in experiments in which no reasonable grouping into blocks suggests itself" — that is, when ignorance prevails. Normally speaking ignorance does not prevail, and real economic and statistical gains can be found by stratifying. Deming (1978) and Tippett (1958) simplified Student's (1911, 1923) proof that stratification (blocking) can reduce sample size requirements by 40% or more, holding variance constant.[27] As Tippett noted, "At the worst" — assuming the rare case that calculated variance between strata is zero — "sampling in strata is no better than random sampling, but it is never worse."

Thus Levitt and List are mistaken: Fisher did not "introduce" blocking in the 1920s and 1930s. Blocking was introduced at least as early as Student (1911) though to repeat one can find a crude version of blocking in, for example, Young (1767) and before that Noah's Ark. Mercer and Hall (1911), Student (1911, 1923), and many other experimentalists before Fisher (notably Beaven (1947)) had been repeating experiments and advocating repeated experiments two decades before Fisher (1925).[28]

But Student took the economic approach further, and put it at the center of his experimental purpose, design, and analysis. In an admittedly crude way, Student suggested and proved in his 1911 article that random layouts are inferior to deliberate balancing in four key variables: precision, efficiency, simplicity, and power to detect a large and real treatment difference when the difference is actually there to detect.[29]

Said Student (1911), "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.

---

stratified sampling over random sampling (Neyman, 1934, pp. 579–585). Neyman credits the idea of "purposive selection" to earlier writers, such as Bowley *et al.*

[27] Deming (1978, pp. 880–881), Tippett (1958, p. 356). In a Riesling vine-and-wine experiment, Meyers *et al.* (2011) used blocking, balancing, and repetition (at $n = 3$ vineyards) to reduce sample size requirements by up to 60%.

[28] Student and Beaven used blocking techniques — the "principle of maximum contiguity" — in repeated experiments as early as 1905 — two decades before Levitt and List claim the techniques were introduced by Fisher: Student (1923, p. 278), Beaven (1947, pp. 237–238).

[29] Student (1938) and Pearson (1938) advanced the proof.

> 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.[30]

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)] *we can judge whether such a difference as we find between the total yields is likely to have arisen from chance.*

[30] Student (1911, 1942, p. 49).

> 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 (1911, 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".[31]

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 Fisher's "introduction" of randomization[32]:

> 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.

Thus long before Fisher entered the scene, Student employed concepts of small sample theory, blocking, paired differences, efficiency, Type II error, and balanced versus random designs of experiments.[33] Indeed, the origin of today's paired t-test can be found in Student (1908, 1911).

---

[31] Student (1911, p. 52); see also: Bruhn and McKenzie (2009), Carson *et al.* (2009), van Es *et al.* (2007).

[32] Student (1911, 1942, p. 51).

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

In 1934 Fisher wrote a letter to Student requesting reprints of Student (1911), Student (1926), and Student (1930). Fisher inquired about the intellectual history of blocking and pairing of differences. Student replied that taking differences of pairs is old, dating back he jokingly said to "Old Noah" and the Ark[34]:

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 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

And yet in published discussions of experimental design Fisher did not credit Student — at all. Prior to Fisher's *Statistical Methods for Research Workers* (1925), Student was frequently credited.[35] Mercer's and Hall's conclusion number "4" deserves emphasis for it reveals a contribution which Student made to experimental thought and to Rothamsted Experimental Station 8 years before Fisher was hired at Rothamsted and 14 years before Fisher began to campaign for completely randomized blocks[36]:

---

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

[35] For discussion, see Neyman (1938); Pearson (1938); Ziliak and McCloskey (2008), Chapters 20–22.

[36] Mercer and Hall, (1911, p. 127).

> (4) For practical purposes the authors [Mercer and Hall] 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.*
>
> This [systematic design] 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.

Student published these findings when the younger Fisher was still a college student at Cambridge.[37]

Levitt and List are correct to assert that "In 1919, Ronald Fisher was hired [at Rothamsted] to bring modern statistical methods to the vast experimental data collected [there]." But they are wrong for at least three reasons when they assert that "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" (Levitt and List, p. 3).

First, the balanced design recommended by Student (1911) and accepted by Mercer and Hall at Rothamsted was not "crude"; it was *more* efficient (Student proved in his tables) than random layouts. Second, as Student (1923) makes clear — and Guinness *Laboratory Reports* confirm: for example, Gosset (1904, 1905b) — Gosset aka "Student" had been repeating experiments in cooperation with large scale farmers and industrialists since 1904, if not earlier (Gosset, 1936). Third, 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 a botched experiment by Fisher and Mackenzie (1923).

Why, then, have Levitt and List credited Fisher and Neyman in the 1920s and 1930s for originating field experiments in economics? The reason seems clear now: randomization. They claim (again without proof) that "randomization was the

---

[37] Ziliak and McCloskey (2008, Chapters 20–22).

lynchpin as the validity of tests of significance stems from randomization theory" (Levitt and List, p. 4).

Yet randomization is not the purpose of an experiment in economics — profit, innovation, quality assurance, and other substantive goals are; in any case, randomization is not the goal and it does not constitute the peak of experimental method in economics. As Student (1923) showed, artificial randomization — however useful when errors are independent and identically distributed (Student, 1908, p. 13) and ignorance prevails — is not as precise as balanced designs are when, as is common, observations and errors are correlated. Maximizing correlation is central to good experimental design. I return to these fundamental issues in Section 6.

5 Neyman Sides with Student: The Comparative Advantage of Balanced Designs

A little digging in the archives suggests that Neyman would reject the authors' claims, too. In a *Journal of the American Statistical Association* obituary notice, Neyman (1938) gave highest honors to the distinguished subject of his notice: William S. Gosset aka Student.[38]

Neyman considered Student's 1923 *Biometrika* article "On Testing Varieties of Cereals" to be the starting point of theory.[39] Nothing of the sort is hinted at by Levitt and List, who try unsuccessfully to align Neyman with Fisher and the randomization school. In the article celebrated by Neyman, Student (1923) pushed his 1911 insights and ambition much further. 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.[40] 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 (Student, 1923). Said Neyman and Pearson (1938, p. 228):

> As a third example of Student's pioneer work I shall quote his first paper[41] on agricultural experimentation, which should be rightly

[38] Contrast Neyman's mainly negative assessment of Fisher's approach: Neyman (1961).
[39] Neyman did not know about Student (1911).
[40] See also: Gosset (1936), Student (1938) and Pearson (1938).
[41] In fact, Student (1911) was Gosset's first published article on the design and evaluation of agricultural field experiments. See *Student's (1931b) Collected Papers* (1942, Pearson and Wishart, Eds.) for the full collection of Student's published work.

> 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.

This was not the first time that Neyman sided in public with Student. At a 1935 meeting of the Royal Statistical Society, Neyman *et al.* (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 *et al.* (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" — that is, the same article by Student (1923) which Neyman considered the start of all theory.

Thus it is surprising to hear from Levitt and List the erroneous claim: "Viewing Neyman's body of work," they say, "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".[42]

Again, this would be a wonderful finding if it were close to the truth. But as Neyman himself explained to Fisher, 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 experimental design with an economic motive.[43]

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

---

[42] Levitt and List, p. 3.
[43] See also: Reid (1982, p. 44).

balancing involves some sacrifice of the normal assumptions. He said in his reply to Fisher[44]:

> As to the Half-Drill-Strip method [Student's preferred balanced layout, sometimes called "ABBA"], 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.[45]

"In a publication already referred to," Neyman (1935, p. 179) said, "I tried to improve the method in this respect."[46]

> 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.

Given these easy-to-find facts about Student, Neyman, and the first wave of field experiments, it is not clear how Levitt and List arrived at such erroneous conclusions about randomization and field experiments.[47]

5.1 Levitt and List Misread Neyman, Fisher and Mackenzie, and Student

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

---

[44] Neyman *et al.* (1935, p. 179).
[45] Student (1923, 1936, 1938) was the first to emphasize this minor weakness of balancing.
[46] Compare: Pearson (1938), and Neyman and Pearson (1938).
[47] Note that Neyman, a great mathematician, did not believe 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," said Neyman, "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 Dabrowska and Speed, 1990).

once — randomization is neither used nor mentioned — in the article by Fisher and Mackenzie.

"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 write.[48] 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.

In 1923 — the same year that Student was comparing random with balanced designs on farm fields across England and Ireland — Fisher had not given much if any thought to the statistical design of experiments, period. 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.

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 (Figure 1) [by Fisher and Mackenzie] illustrates [Yates said] how little attention was given to matters of layout at that time.[49] It is indeed difficult to see how the arrangement of the varietal plots [designed by Fisher and Mackenzie] was arrived at.

Thus Fisher's design in 1923 was neither randomized nor balanced: "the arrangements for the farmyard manure and no farmyard manure blocks are almost but

---

[48] The correct spelling is "Mackenzie" (Fisher and Mackenzie, 1923).
[49] But compare Student (1911, 1923), two seminal articles omitted by Yates.

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 want their readers to believe otherwise.

Yates continued (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.

Yet to understand the future development and popularity of randomization, it is important to understand the immediate personal and social ramification of Fisher's spoiled experiment. In a letter of correspondence, 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 article by Fisher and Mackenzie. Student replied to Fisher in a letter of July 1923, commenting on the Fisher and Mackenzie experiment 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."[50]

Ouch. That must have hurt Fisher, whose career in 1923 was still getting started at Rothamsted.

Still, to his credit, Fisher replied to Student's helpful if embarrassing letter, asking Student what he, Student, would do differently to the design of the experiment.[51] 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

---

[50] Letter #29, July 25, 1923, in Gosset, 1962.
[51] Letter #s 20, 23, in Gosset (1962).

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. "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)].[52]

Fisher did not warm 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 and field experiments. Student and others have priority and so the first main claim of their article — the first wave thesis — is strongly rejected. Neyman sided with Student, rejecting randomization in favor of balanced designs. And Fisher and Mackenzie (1923) — a famously spoiled experiment — does not mention randomization let alone use it. Levitt and List (2009) have misidentified the canon.

5.2 Fisher's Revolt: The False Sociology of Randomization, 1923 to 1925

How, then, did it happen that randomization appears as a "principle" of design in Fisher (1925) — less than two years after Student's critical letters?

Student's second letter to Fisher must have put a bee in a bonnet.[53] Regardless, between August 1923 and October 1924, Fisher suddenly became a radical proponent of artificial randomization in the design of experiments — the very design

---

[52] Letter #s 29, 30, in Gosset (1962); Gosset (1923), quoted in Pearson (1990, p. 58).

[53] 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. Fisher 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]); for discussion see: Ziliak (2011b) and Ziliak and McCloskey (2008, chapters 20–24).

which his behind-the-scenes muse and mentor, Student (1911, 1923), had disproved and warned against (for five published volumes of Student-Fisher correspondence see, Gosset, 1962). 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 Student? Perhaps (see: Gosset, 1962). 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 (at Fisher's request) the basic principles of experimental design.[54] 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 since 1905 by Student, Beaven, and others. 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, containing 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."[55] "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's objections; 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.

---

[54] Elsewhere I have shown evidence of Fisher's growing intolerance for Student's, Jeffreys', Neyman's, Pearson's, Deming's, Wald's, Savage's, and others' economic approach to the logic of uncertainty: Ziliak (2010), Ziliak and McCloskey (2008, Chapters 20–22).

[55] Quoted in Pearson (1990, p. 59).

The article by Levitt and List, however flawed, does raise an interesting question for the history of experimental economics: when was artificial randomization introduced to statistically based experiments in economics?

Normally people answer as Cochrane, Yates, and now Levitt and List do, crediting Fisher. Yates (1964), who for many years worked side by side with Fisher, and coauthored 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 — in fact, 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 long before Fisher and Neyman, in a seminal Monte Carlo simulation Student (1908) describes in his *Biometrika* article "On the Probable Error of a Mean"[56]:

> Before I had succeeded in solving my problem analytically [that is, Student's z distribution, as Student's t was first called], 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.

---

[56] I am not claiming that Student (1908) is the first use of artificial randomization in a formal statistical testing 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 distribution for a profit-seeking firm (Guinness) is a seminal use of artificial randomization in experimental *economics*.

So Student was not completely anti-randomization; he used randomization to simulate Student's distribution when ignorance prevailed and observations were independently distributed — just as Jeffreys, Neyman, Deming, Cox, Tippett, and others suggested after him. The primary motive for Student's 1923 article was to demonstrate to statisticians, farmers, and the Managing Director of Guinness the economic and statistical advantages and disadvantages of balanced versus random layouts of competing barley varieties in a small number of repeated experiments (see also: Gosset,1936).

Levitt and List cite Student (1923) but they fail to mention the content, purpose, and main finding of Student's work, which is against randomization and in favor of balancing. They omit altogether his massive evidence from comparative tests in a private-seeking environment.

Unlike Fisher's method, Student's experimental method was substance and profit-seeking throughout, from the choice of sample size on up, raising the cost of neglecting Student further still.[57] For example, the profit-conscious Student (1923, p. 281) compared 193 plots on 18 farms for eight varieties of barley grown in different barley growing regions of Ireland:

> *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" calculated using Student's and Fisher's ANOVA structure, the Guinness scientist concluded[58]:

> 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 an improvement over chessboards and randomization which he discovered in collaboration with Beaven (1947): "The Half-Drill Strip Method," also known as "ABBA".[59]

---

[57] Ziliak (2008), Ziliak and McCloskey (2008).
[58] Student (1923, p. 282).
[59] Student (1938, pp. 364–378).

6 Student Against the Randomization Thesis

"We now proceed to the most accurate method yet devised for field trials" (Student, 1923).[60] What makes the half-drill strip or ABBA method so effective compared to random and the other balanced designs (chessboards, for example)? In general if one or more explanatory variables are known to have a systematic spatial or temporal correlation with respect to the output of interest (the dependent variable) the allocation of treatments and controls 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. Balancing can be done in a variety of ways but the essence of the idea is to give equal exposure of treatments and controls to both systematic and random sources of fluctuation. The advantage of balancing has been shown not only in medical, epidemiological, and pharmaceutical experiments but also in social policy and development economics.[61]

Early examples of balanced designs in crop yield trials are chessboard and Knight's move. Take the Knight's move, for example, balancing blocks in an 8×8 chessboard-type layout.[62] Suppose the brewer is comparing yields of eight different varieties of barley sown in the field — as in 1923 Student did — the eight varieties matching the 8 rows and columns of an actual chessboard. How shall the seeds be planted according to the design of the experiment? Knight's move says to balance the distribution of systematic sources of fluctuation (in this case, the diminishing marginal productivity of the soil) 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 blocks up and one over from another 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 $n \times k$ design and the number of experiments (farms) in the series.

Consider a simpler case: a single comparison, testing (in a yield trial, say) variety A against variety B.

Assuming 8×8 blocks, the experimental field contains $n = 64$ blocks in which to 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

---

[60] Student (1923, p. 285).
[61] Altman *et al.* (2001), Rothman *et al.* (2008), Heckman (1991), van Es *et al.* (2007), Bruhn and McKenzie (2009), Ziliak (2010), Harrison (2011).
[62] This section is borrowed from Ziliak (2011b, pp. 14–24).

seeds from both A and B. I assume here that each block gets a unique "treatment".) The problem with random assignment of A and B to blocks is that by chance the recommended allocation might be:

$$\begin{array}{l} \text{A A A A A A B B} \\ \text{A A A A A A B A} \\ \ldots \text{ etc.} \\ \longrightarrow \quad \text{Direction of increase in soil fertility (higher yielding soil)} \end{array} \tag{1}$$

Another random draw could produce blocks of this sort:

$$\begin{array}{l} \text{B B B B B A A B} \\ \text{B B B B B A A A} \\ \ldots \text{ etc.} \\ \longrightarrow \quad \text{Direction of increase in soil fertility (higher yielding soil)} \end{array} \tag{2}$$

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 (1) and (2) — though random — produce biased mean squared errors and parameter estimates with respect to a major source of fluctuation — differential soil fertility. In example (1) the As are bunched up and growing in the very worst soil; thus the yield of the B's will be artificially high, and the real treatment difference, A − B, will be undetermined.[63]

Student and collaborators found again and again, in repeated trials, that deliberate balancing — though adding to the "apparent" error, that is, to Type I error in ANOVA terms, actually *reduces* the real error of the experiment, minimizing Type II error and errors from fixed effects, such as non-random soil heterogeneity.[64]

Examples (1) and (2) 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. Random layouts analyzed with Student's test of significance will

---

[63] In other words, complete randomization raises the risk of Simpson's Paradox, leading to a false reversal of actual treatment effect. Lindley (1991, pp. 47–48) demonstrates that stratification and balance are the best available means for avoiding Simpson's Paradox.

[64] Student (1938, pp. 364–372).

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 became (grammarians have to admit) perfectly balanced. The ABBA layout is:

$$\begin{array}{l} \text{A B B A A B B A} \\ \text{A B B A A B B A} \\ \text{A B B A A B B A} \\ \ldots \text{ etc.} \end{array} \tag{3}$$

One virtue of the ABBA design is that it minimizes bias caused by differential soil fertility. Given the built-in symmetry of ABBA, 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, except by accident, 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 (1) and (2) above and it is 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 details). 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."[65] 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). In

[65] Pearson (1938, pp. 163–164); Pearson shows how to adjust ANOVA and Student's test of significance to accommodate the ABBA structure.

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 — he said to Fisher[66]:

> 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.

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).

7 The Power and Efficiency of Balanced Designs

The intuition behind the higher power of ABBA and other balanced designs to detect a large and real treatment difference was given by Student in 1911.[67] "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".[68] He used this "principle

[66] Student (1938, p. 366).
[67] See also: Beaven (1947, pp. 273–275).
[68] Student (1911, p. 128).

of maximum contiguity" often, for example when he illustrated the higher precision and lower costs that would be 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).[69]

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 A's and B's.

The idea is that of stratification: to create homogeneous growth conditions and to maximize informational content per unit of cost. (In a study of entrepreneurship, for example, imagine studying random samples of pairs or twins, so to speak, such as Ben and Jerry, in an experiment with other pairs both like and unlike Ben and Jerry.)

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 giant of statistics not mentioned by the authors — clarified Student's theory.[70] 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:

$$x_{\mathrm{A}} = m_{\mathrm{A}} + \delta_{\mathrm{A}}$$

---

[69] Student (1923 [1942], p. 95). 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." Likewise, 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, neither random nor balanced.

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

and

$$x_{\mathrm{B}} = m_{\mathrm{B}} + \delta_{\mathrm{B}}$$

and thus:

$$x_{\mathrm{A}} - x_{\mathrm{B}} = (m_{\mathrm{A}} - m_{\mathrm{B}}) + \delta_{\mathrm{A}} - \delta_{\mathrm{B}} = (m_{\mathrm{A}} - m_{\mathrm{B}}) + \Delta_{\mathrm{AB}} \tag{4}$$

where x_i is the yield from the ith block or plot, m_i is the yield in the ith 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.[71]

Students of Heckman and Zellner, for example, will not be surprised by what follows from Student's and Pearson's set up, which strives for real error minimization (not mere statistical significance). 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 $\Delta_{\mathrm{AB's}}$ — the real treatment effects (or causes if you will) on yield, within and between farm fields.[72]

> 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.[73] Each 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.

Using a simple workhorse formula, Student argued that the ABBA layout 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

---

[71] Pearson (1938, pp. 163–164).

[72] Gosset (1936, p. 118).

[73] 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, p. xiv). When Beaven died (in 1941) Guinness acquired ownership and continued to produce at the 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.

he used to measure the variance of mean differences, A − B, he got in 1905 from Karl Pearson, during a July visit to Pearson's summer house:

$$\sigma^2_{\mathrm{A-B}} = \sigma^2_{\mathrm{A}} + \sigma^2_{\mathrm{B}} - 2\rho_{\mathrm{AB}}\sigma_{\mathrm{A}}\sigma_{\mathrm{B}} \tag{5}$$

where σ^2 is variance and ρ is the Pearson correlation coefficient.[74]

Unlike randomization, 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 private sector firm — such as Guinness's large scale brewery — wants when precision and profit are ultimate goals.

Fisher's randomization — and thus most of the current field experiments in economics — ignore the fundamental importance of the correlation coefficient, ρ; assuming independent and identically distributed observations in imaginary replications, artificial randomization seeks only to minimize σ^2_{A} and σ^2_{B}. 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 noted], so the ordinary formulae for combining errors of observation which are based on randomness are even less applicable than usual.

How unexpected, then, to discover Levitt and List (2009, p. 4) asserting 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

---

[74] K. Pearson, quoted by Gosset (1905a,b), Guinness Archives; reprinted: Pearson (1939, p. 212).

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, 1908, 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, 1939, p. 242). "Fisher proceeds . . . to *make it* into a random error" (p. 243; italics in original). But[75]:

> 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.

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 again, citing Student (1923, 1938) as the source of his ideas[76]:

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

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,

---

[75] Jeffreys (1939, p. 243)

[76] Jeffreys (1939, p. 243); Jeffreys believed that Student's methods were in basic agreement with his own Bayesian approach (Jeffreys, 1939, pp. 379, 393, 369–400). Ziliak (2008) and Ziliak and McCloskey (2008) describe Student's early endorsement of Bayes's theorem in practical work.

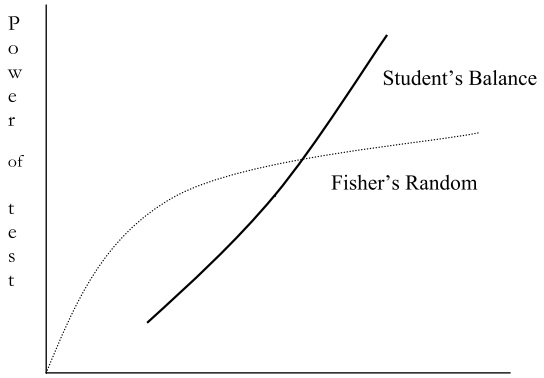

Figure 1. Size of real treatment difference measured by the variance of Student's Δ's.

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).

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:[77]

> 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.[78]

---

[77] Gosset (1905a); reprinted in Pearson (1939).

[78] Contrast the profit-seeking purpose of experiments according to Gosset, Pearson, Jeffreys, and Savage with the anti-economic approach taken by 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".

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.

8 The Third Wave Thesis and Three Easy Ways to Improve Field Experiments in Economics

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. The general point that Student proved and Fisher ignored is that artificial randomization gives less assurance in results, not more, when observations are spatially and/or temporally correlated and confounded by one or more variables.

An important general result for experimentation 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" (Levitt and List, p. 14) offering no evidence or explanation for their belief.

Typically the goal of the economist working for a government body or private sector firm is to discover profitable or other practical differences at minimum cost. This was the central point of Student's (1931a) proposal to save the Health Department of Scotland tens of thousands of pounds, and increase precision, by studying $n = 50$ pairs of biological twins rather than the $n = 20{,}000$ children actually studied in the biased and expensive Lanarkshire Milk Experiment. Maximizing net benefit is, likewise, a central point of Harrison's (2011) important survey of field experiments in development economics.

In the real sector of the economy — where Student worked — experimentation is not an academic problem. It is a business problem, leading to a decision. 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.[79] 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.

Yet as Beaven (1947, p. 293) noted long ago, "Many 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".[80] Unfortunately then the authors' third claim — the third wave thesis — must also give way: the authors are not the first experimentalists in history to run randomized experiments with private sector firms. More than a century ago, Gosset (1904) combined laboratory and field experiments for the Guinness Brewery. Despite their many valuable contributions to economics since the mid-1990s it is premature to claim the authors have seen "much deeper" than Student, Jeffreys, Neyman, Pearson and others from "previous generations".[81]

What are three easy ways to increase the value of field experiments in economics? The experimental economist can innovate while increasing precision and decreasing costs for the firm that can:

1. *Demonstrate in a small number of repeated experiments, using the smallest sample sizes possible, probable ability to achieve a minimally profitable bottom line result.*

---

[79] "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.

[80] An anonymous referee questions 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 would be shared by Student and Savage, for example, and Heckman now. 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 the distinction between observation and experiment may be in ordinary practice, I do not yet see that it admits of any solid analysis".

[81] Levitt and List, p. 15.

For example, Gosset (1904) showed the Guinness Board that he could mix just $n = 4$ malt extracts to get better than 10-to-1 odds of being within 0.5 degrees saccharine of the actual Guinness standard, which was 133 degrees saccharine per barrel of malt, to maintain a stable level of alcohol content and to minimize the tax paid on alcohol (as described by Ziliak, 2008). Contrast Karlan and List (2007, p. 1777), who report a small rate of return given the size of their costly, unbalanced, and unrepeated experiment on $n = 50{,}083$ public radio donors;

2. *Design and repeat a small series of independent experiments (or samples) stratified and balanced in covariates with respect to both non-random and random sources of error.*

Student's principle of maximum balance in design is, when combined with the economic interpretation of uncertainty, well suited to the purpose of economic experiments. Fisher's bright line rules about randomization and significance are not; Fisher's rules neglect both cost and benefit, causing — among other things — needlessly large sample sizes. Still, as Student, Neyman, Deming, Cox and others have observed, a small random sample is sometimes useful at a preliminary stage of investigation. A small randomized trial or sample survey may help when ignorance of data, of strata, and of non-random confounding factors prevail (differential soil fertility, or, to take an example from human soil, differential beliefs and motivation). To maximize information per unit of cost, the experimentalist might examine as molecular chemists do, small random samples from near and widely dispersed units. (Exploratory data analysis and visualization — what Savage called "interocular trauma" — helps at this stage.) Sorting random data into homogeneous units is in any case, as Deming notes, basic to scientific discovery. The data must be reduced. Thus discovery of relevant strata appears to be an essential feature of the information cost-minimization problem, the problem which the experimental economist and firms seek to solve.[82] Holding cost constant, artificial randomization does not produce as much knowledge, information, or profit compared to stratified and balanced designs using the principle of maximum contiguity;

3. *Replace — as the U.S. Supreme Court already has — bright-line rules of statistical significance (such as $p < 0.05$) with alternative scenarios for profitable and losing odds.*

---

[82] An outstanding recent example is Meyers *et al.* (2011) which, after balancing and blocking, reduced sample size requirements in a multi-vineyard, vine-planting experiment on Rieslings by up to 60%.

Finally, field experiments in economics could become more valuable if individual investigators would form rational economic gambles and judgments about the posterior probability of observing some magnitude or range of magnitudes of economic significance from an experiment or series of experiments, due attention being paid to the opportunity cost of the experiment itself.[83]

The null hypothesis test procedure promoted by Fisher, using a maximum likelihood function and test of significance without a prior probability and loss function, is a staple of the literature. Yet the increasingly familiar procedure of randomization plus statistical significance (the yes/no, exists/does not exist world of significance testing at the 0.05 level) does not encourage rational economic gambling or judgment. The usual procedure — by insisting on 19-to-1 odds or better — turns a quantitative/economic problem into a qualitative/either-or decision, reducing efficiency and distorting judgment (Ziliak and McCloskey, 2008).[84] But as Jeffreys said, echoing Student, we want to know the probability of some "minimum difference" between the treatment effect and the industry standard — the bottom line of the experiment must be asserted. Savage (1954, p. 116) wrote in tacit agreement with Student, Jeffreys, Guinness stakeholders, and economic logic: "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".

Can completely randomized blocks, judged by a fixed and arbitrary rule of statistical significance, provide these or better results? No. Can Student's balanced designs and small sample economic approach to the logic of uncertainty help? It seems so.

Refutation of results from completely randomized trials is the scientific finding again and again, in economics and agriculture as much as in case-controlled epidemiology and medicine, Heckman, Vytlacil, Deaton, Altman, Rothman, Ziliak and others have shown after Student. Balanced designs are more powerful and efficient in discrete choice models simulated by Carson *et al.* (2009) in agricultural economics

---

[83] For examples, see McCloskey and Ziliak (2010). See also the valuable if underutilized textbooks by Lindley (1991) and Raiffa and Schlaifer (2000).

[84] If $p = 0.05$ there is — assuming good experimental design, controls for confounding, ample repetition, and other things equal — a 0.95/0.05 or 19-to-1 likelihood or better that an observed difference is not randomly different from the null, that there is some real effect. But the p-value does not answer the how much question, and cannot. And 2-to-1 odds might be enough to accept a gamble on a brain tumor surgery, construction project, or horse race. See Ziliak (2011a) and McCloskey and Ziliak (2010) to read more about the value of rational gambles made at less than 19-to-1 odds in the context of a major case of securities law decided by the U.S. Supreme Court, Matrixx v. Siracusano, March 22, 2011.

and, in development economics, in work by Bruhn and McKenzie (2009), comparing the relative performance of balanced and random designs of experiments related to culture, gender, education, income, and the World Bank.[85]

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). Student's balanced experiments were plenty random, causing no problem for Student's test of significance; the Irish barley experiments, sponsored by Guinness and the Irish Department of Agriculture, were conducted on scattered farms in 10 or 12 different barley-growing districts of Ireland, random enough.

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 (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.

References

Altman, D. G., K. F. Schultz, D. Moher, M. Egger, F. Davidoff, D. Elbourne, P. Gotzsche, and T. Lang 2001. "The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration." *Annals of Internal Medicine* 134: 663–691.

Banerjee, A. and E. Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way To Fight Global Poverty.* New York: Public Affairs.

Beaven, E. S. 1947. *Barley: Fifty Years of Observation and Experiment.* London: Duckworth.

Berk, R. A. and D. A. Freedman. 2003. "Statistical Assumptions as Empirical Commitments." In *Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger.* T. G. Blomberg and S. Cohen, (eds.), Aldine de Gruyter, pp. 235–254.

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

Bruhn, M. and D. McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Economics." *American Economic Journal: Applied Economics* 1: 200–232.

[85] An important paper by Bruhn and McKenzie (2009) compares the power of balanced and random designs using methods pioneered by Student (1923, 1938) and Pearson (1938). Ironically, Bruhn and McKenzie cite Fisher but not Student and Pearson.

Carson, R. T., J. J. Louviere, and N. Wasi. 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 *On the History of Statistics and Probability*, D. B. Owen, (ed.), New York: Marcel Dekker Inc., p. 126.
Cochrane, W. G. 1989. "Fisher and the Analysis of Variance." In *R. A. Fisher: An Appreciation*, New York: Springer-Verlag, pp. 17–34.
Concise Dictionary of National Biography, Part II, 1901–1950. Oxford: Oxford University Press.
Cox, D. 1958. *Planning of Experiments*. New York: Wiley.
Deaton, A. 2007. "Evidence-based Aid Must not Become the Latest in a Long String of Development Fads." In *Making Aid Work*, A. Banerjee, (ed.), Cambridge: MIT Press, pp. 60–61.
Deming, W. E. 1978. "Sample Surveys: The Field." In *International Encyclopedia of Statistics*, W. H. Kruskal and J. M. Tanur, (eds.), New York and London: The Free Press (Macmillan), pp. 867–884.
Duflo, E., R. Glennerster, and M. Kremer. 2007. *Using Randomization in Development Economics Research: A Toolkit*. MIT Department of Ecconomics and J-PAL Poverty Action Lab.
Es van, H.M., C. P. Gomes, M. Sellman, and C. L. van Es. 2007. "Spatially-Balanced Complete Block Designs for Field Experiments." *Geoderma 2007* 140: 346–352.
Fisher, R. A. 1925 [1928]. *Statistical Methods for Research Workers*. New York: G.E. Stechart.
Fisher, R. A. 1926. "Arrangement of Field Experiments." *Journal of Ministry of Agriculture* 33: 503–513.
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." Edinburgh: Oliver & Boyd.
Fisher, R. A. and W. A. Mackenzie. 1923. "Studies in Crop Variation: II. The Manurial Response of Different Potato Varieties." *Journal of Agricultural Science* 13: 311–320.
Fisher, R. A. and F. Yates. 1938. *Statistical Tables for Biological, Agricultural and Medical Research.*, Sixth edition, Edinburgh: Oliver and Boyd.
Gosset, W. S. 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: 115–136.
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*. New York: E.P. Dutton and Company.
Harrison, G. W. 2011. "Randomization and Its Discontents." *Journal of African Economies* 20: 626–652.
Harrison, G. W. and J. A. List. 2004. "Field Experiments." *Journal of Economic Literature* 44: 1009–1055.

Heckman, J. J. 1991. *Randomization and Social Policy Evaluation*. NBER Working Paper No. T0107. Cambridge: National Bureau of Economic Research.
Heckman, J. J. and J. Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives* 9: 85–110.
Heckman, J. J. and E. J. Vytlacil. 2007. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." In *Handbook of Econometrics 6B*, J. J. Heckman and E. Leamer, (eds.), Amsterdam: Elsevier, pp. 4779–4874.
Herberich, D. H., S. D. Levitt, and J. A. List. 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.)*. Chicago: University of Chicago Press, p. 220.
Jeffreys, H. 1939. *Theory of Probability*. London: Oxford University Press. Third revised edition.
J-PAL Poverty Action Lab. 2010. *When Did Randomized Evaluations Begin*? Cambridge: Poverty Action Lab, MIT. http://www.povertyactionlab.org/methodology/when/when-did-randomized-evaluations-begin
Karlan, D. and J. List. 2007. "Does Price Matter in Charitable Giving? Evidence From a Large-Scale Natural Field Experiment." *American Economic Review* 97: 1774–1793.
Karlan, D. and J. Appel. 2011. "More Than Good Intentions: How a New Economics is Helping to Solve Global Poverty." New York: Dutton.
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: 1019–1030.
Lanham, R. A. 1991. *A Handlist of Rhetorical Terms*. Los Angeles: University of California Press.
Leon, A. C., H. Demirtas, and D. Hedeker. 2007. "Bias Reduction with an Adjustment for Participants' Intent to Drop Out of a Randomized Controlled Clinical Trial." *Clinical Trials* 4: 540–547.
Leonard, A. 2009. "Celebrate the History of Statistics: Drink a Guinness. How a Master Brewer Forged New Ground in the Quantitative Progress of Science." *Salon*: September 28. http://mobile.salon.com/tech/htww/2009/09/28/guinessometrics/index.html
Levitt, S. D. and J. A. List. 2009. "Field Experiments in Economics: The Past, the Present, and the Future." *European Economic Review* 53: 1–18.
Lindley, D. 1991. *Making Decisions*. New York: Wiley.
List, J. A. 2009. "An Introduction to Field Experiments in Economics." *Journal of Economic Behavior and Organization* 70: 439–442.
List, J. A. and J. Schogren. 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. and S. T. Ziliak. 2009. "The Unreasonable Ineffectiveness of Fisherian 'Tests' in Biology, and Especially in Medicine." *Biological Theory* 4: 44–53.
McCloskey, D. N. and S. T. Ziliak. 2010. Brief of *amici curiae* by statistics experts Professors Deirdre N. McCloskey and Stephen T. Ziliak in support of respondents, Vol. 09-1156, Supreme Court of the United States, Washington DC. Edward Labaton *et al.* Counsel of Record, (ed.), Matrixx *et al.* vs. Siracusano and NECA-IBEW Pension Fund, filed Nov. 12, 2010.
Mercer, W. B. and A. D. Hall. 1911. "The Experimental Error of Yield Trials." *Journal of Agricultural Science* 4: 107–127.
Meyers, J., G. Sacks, H. van Es, and J. Vanden Heuvel. 2011. "Improving Vineyard Sampling Efficiency via Dynamic Spatially Explicit Optimisation." *Australian Journal of Grape and Wine Research* 17: 306–315.

Moore, D. and G. McCabe. 1998. *Introduction to the Practice of Statistics*. Third edition. New York: W. H. Freeman.
Neyman, J. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97: 558–625.
Neyman, J. [See below: Splawa-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: 145–154.
Neyman, J., K. Iwaszkiewicz, and S. Kolodziejczyk. 1935. "Statistical Problems in Agricultural Experimentation." *Supplement to the Journal of the Royal Statistical Society* 2: 107–180.
Neyman, J. and E. S. Pearson. 1938. "Note on Some Points on 'Student's' Paper on 'Comparison Between Balanced and Random Arrangements of Field Plots." *Biometrika* 29: 379–388.
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: 210–250.
Pearson, E. S. 1990 [posthumous]. *'Student': A Statistical Biography of William Sealy Gosset*. Oxford: Clarendon Press. Edited and augmented by R. L. Plackett, with the assistance of G. A. Barnard.
Peirce, C. S. and J. Jastrow. 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*. New York: Wiley.
Raiffa, H. and R. Schlaifer. 2000. *Advanced Statistical Decision Theory*. New York: Wiley.
Reid, C. 1982 [1998]. *Neyman: A Life*. New York: Springer.
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., S. Greenland, and T. Lash. 2008. *Modern Epidemiology*. Philadelphia: Lippincott, Williams & Wilkins.
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*. New York: Dover.
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: Oxford University Press, 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 D. M. Dabrowska and T. P. Speed, (eds.).
Stigler, G. J. 1969 [1982]. "Does Economics Have a Useful Past?" *History of Political Economy* 2: 217–230. In *The Economist as Preacher*. G. J. Stigler, (ed.), Chicago: University of Chicago Press, pp. 107–118.
Stigler, S. M. 1986. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge: Harvard University Press.
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: 351–360.

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 Reprinted in: E. S. Pearson and J. Wishart, (eds.) (1942). Student's Collected Papers. London: University College of London, pp. 49–52.
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 and J. Wishart, (eds.) (1942). Students Collected Papers. London: University College of London, pp. 121–134.
Student. 1927. "Errors of Routine Analysis." *Biometrika* 19: 151–164.
Student. 1930. "Evolution by Selection." *Eugenics Review* 24: 293–296.
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: E. S. Pearson and J. Wishart, (eds.) (1942). *Student's Collected Papers*. London: University College 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: 363–378.
Student. 1942 [posthumous]. *Student's Collected Papers*. London: University College London. E. S. Pearson and J. Wishart, (eds.).
Taleb, N. N. 2005. *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*. New York: Random House.
Tippett, L. 1958. *The Methods of Experiments*. New York: Wiley.
Varian, H. 2011. "Are Randomized Trials the Future of Economics? Federalism Offers Opportunities for Casual [sic] Experimentation." *The Economist*: 27th April. http://www.economist.com/node/21256696
Wood, T. B. and F. J. M. Stratton. 1910. "The Interpretation of Experimental Results." *Journal of Agricultural Science* 3: 417–440.
Yates, F. 1964. "Sir Ronald Fisher and the Design of Experiments." *Biometrics* 20: 307–321.
Young, A. 1767. *The Farmer's Letters to the People of England*. London: W. Nicholl.
Zellner, A. 2004. *Statistics, Econometrics, and Forecasting*. Cambridge: Cambridge University Press.
Zellner, A. and P. Rossi. 1986. "Evaluating the Methodology of Social Experiments." *Proceedings of the Federal Reserve Bank of Boston*, pp. 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. "Matrixx vs. Siracusano and Student vs. Fisher: Statistical Significance on Trial." *Significance* 8: 131–134.
Ziliak, S. T. 2011b. "W.S. Gosset and Some Neglected Concepts in Experimental Statistics: Guinnessometrics II." *Journal of Wine Economics* 6: 252–277.
Ziliak, S. T. and D. N. McCloskey. 2008. *The Cult of Statistical Significance: How the Standard Error Costs us Jobs, Justice, and Lives*. Ann Arbor: University of Michigan Press.