The Unprincipled Randomization Principle in Economics and Medicine

By Stephen T. Ziliak and Edward R. Teather-Posadas

March 29, 2014

Department of Economics
Roosevelt University
430 S. Michigan Ave
Chicago, IL 60605 USA

Email: sziliak@roosevelt.edu
Email: eteather@roosevelt.edu

Forthcoming in *Oxford University Press Handbook on Professional Economic Ethics*
(New York: Oxford University Press, 2014), Eds. George DeMartino and Deirdre N. McCloskey

Abstract: Over the past decade randomized field experiments have gained prominence in the toolbox of economics and policy making. Yet randomization enthusiasts have paid little attention to the ethical issues, economic costs, and theoretical difficulties caused by the so called randomization principle. Randomized trials give placebos or no treatment at all to vulnerable individuals, withholding best treatments from the control group. Randomization has been proven to be less precise and less efficient than "Student's" balanced alternatives - particularly when effect sizes and confounding from unobserved systematic effects are large. From medicine to economics, randomized trials are rarely if ever repeated. A good thing, perhaps, given that randomizers are willing to sacrifice the well-being of study participants in order to "learn". We consider here the ethics and logic of that sacrifice. We present new evidence from a 25 question survey of randomization, statistical significance, and validity we applied to all the full length articles using randomization techniques and published in the *American Economic Review*, 2000-2009 and *New England Journal of Medicine*, 2000-2002. A study of history shows that today's "principle" of randomization was fabricated out of nothing by R.A. Fisher, in a little known battle he waged in the 1920s against W.S. Gosset aka "Student". We conclude that the most reliable ethical character of economics, Adam Smith's "impartial spectator," would not approve of randomized trials as practiced in economics and medicine.

Keywords: randomization, balance, ethics, Simpson's paradox, field experiments, W.S. Gosset

JEL codes: A11, B23, C9, C93, O2

The Unprincipled Randomization Principle in Economics and Medicine

By Stephen T. Ziliak and Edward R. Teather-Posadas

Acknowledgments We thank the editors, George DeMartino and Deirdre N. McCloskey, and Tim Harford, Jason Houle, Gary Langer, Casey Mulligan, Martin Ravallion, Flora S. Ziliak, and James P. Ziliak for comments on a previous version of this manuscript. Our thanks also to Julian Alston and Koen Deconinck for comments made at the 6th annual conference of the American Association of Wine Economists, Princeton, NJ, June 2012.

Over the past decade randomized field experiments have gained prominence in the toolbox of economics and policy making. Yet enthusiasts for randomization have perhaps paid not enough attention to conceptual and ethical errors caused by complete randomization.

Many but by no means all of the randomized experiments are being conducted by economists on poor people in developing nations. The larger objective of the new development economics is to use randomized controlled trials to learn about behavior in the field and to eradicate poverty.[1] The number of randomized trials in economics is increasing in the United States and Europe, too (Levitt and List, 2009). Herberich, Levitt and List (2009) is representative of the work being done. They believe that a whole hog return of randomized trials to agriculture - including crop yield and variety trials - could "return" to "glory" (p. 1259) a field they take to be all dried up (but see, for example, Ziliak [2014, 2011b] and Meyers et al. [2011]).[2]

Not everyone admires randomized trials in economics, foreign or domestic. Angus Deaton (2007) calls the leaders of the randomization movement in development economics *randomistas*. Deaton does not believe that randomization of treatments and controls can solve the fundamental identification problem of econometric parameter estimation, and he – like William Easterly (2009), Dani Rodrik (2008), Sanjay Reddy (2012) and others - questions the macroeconomic validity of their one-off, small market experiments. "Randomization is a metaphor and not a gold standard," Heckman and Vytlacil (2007, p. 4836) have noted. "Student's" collaborator, the experimental maltster and barley farmer, Edwin S. Beaven, observed long ago (1947, p. 293) in a reply to Ronald A. Fisher (1925, 1935), "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".

Mr. Beaven, we believe, was right. Randomization is not the purpose of an experiment in business or economics. Profit is. Or quality assurance is. Or growth is. Randomization is not, and in fact, as one of us has shown, randomized field experiments were tried and rejected in an economic context more than a century ago, by the pioneer of randomized trials in economics.[3] Between 1904 and 1937, William Sealy Gosset aka Student (1876-1937) – the same "Student" of

[1] For example, Banerjee and Duflo (2011), Duflo, Glennerster, and Kremer (2007), and Karlan and Appel (2011).

[2] Herberich, Levitt, and List do not seem to realize that randomized field experiments began in agricultural economics, starting with Student's (1911, 1923) rejections of them in favor of deliberate and balanced designs.

[3] Ziliak (2014, 2011b, 2010a, 2008); Ziliak and McCloskey (2008, chps. 20-22).

Student's test of statistical significance – designed a large number of barley yield and variety experiments for his employer, Arthur Guinness & Sons Ltd., Dublin, comparing the relative performance of random and balanced designs, in order to advance the beer and Guinness bottom line.[4]

Balanced designs are deliberate, systematic allocations of treatments and controls to the experimental units. Student (1911, 1923) discovered the advantage of balancing allocations symmetrically with respect to relevant strata or blocks (in drug trials, for example, important strata include bodyweight and age). He also discovered the advantage of balancing allocation of treatments with respect to non-random fixed effects (observed and unobserved) which are known in advance to spoil randomized experiments by creating a source of uncontrolled heterogeneity and variance in experimental units. For example, in crop yield trials the confounding variable is differential soil fertility - a big source of variance in the object of interest. Student's balanced designs defeated the completely randomized whenever the experimental result really mattered, that is, whenever economic differences between treatments were high enough for the farmer and Guinness brewer to care. Beaven, Jerzy Neyman, Egon Pearson, and Sir Harold Jeffreys, among others, admitted the superiority of Student's balanced designs (see also: Heckman and Vytlacil, 2007). Given the history and outcome of the Student-Fisher debates on randomization, the trouble with RCTs in medicine and pharmacology (Altman et al., 2001), and the high cost of producing what are after all ambiguous results from large, unbalanced, and unrepeated experiments, we find - the loss of experimental knowledge after Student is non-trivial.

There are, in addition, prudential and other ethical implications of a practice that deliberately withholds already-known-to-be best practice treatments from one or more human subjects. Randomized trials often give nil placebo or no treatment at all to vulnerable individuals, withholding (in the name of science) best treatments from the control group.

We present new evidence from a 25 question survey of randomization, statistical significance, and validity which was applied to all the full length articles employing randomization techniques and published by the *American Economic Review*, 2000-2009, and, for comparative purposes, the *New England Journal of Medicine*, 2000-2002.

George DeMartino (2011) notes that economists need a handbook of ethics (such as this one) because economists – and the methods and policies we promote - affect human lives. The use of field experiments in economics and medicine is not itself objectionable. The objection is that no principle can be discerned behind randomization.

I. Blinded Me With Science: The Chinese Children's Eyesight Experiment

A recent example of a randomized and controlled field experiment in both economics and medicine is "Visualizing Development: Eyeglasses and Academic Performance in Rural Primary Schools in China" (Glewwe, Park, and Zhao, 2012).

The development economists wished to know if wearing corrective eyeglasses might enable sight-defective kids to perform better at school. The economists were inclined to believe "yes," that wearing eyeglasses would help. In contemporary medical ethics one would say there is no basis for a randomized trial in this instance because "clinical equipoise" does not exist, meaning that the scientific community is not indifferent between treating and not treating sight

[4] See Student's (1942) collected papers, edited by Egon S. Pearson, and especially therein Student (1938, 1923, 1911) and Gosset (1936).

defective people with eyeglasses (Howick, 2011; Altman et al., 2001; Freedman, 1987). But too many randomized trials are, we find in our survey, violating even this simple norm of conduct.[5] Although many thousands of schoolchildren from rural China were discovered on investigation to be sight defective and lacking in corrective glasses at the time the experiment began, not all of those children were provided with glasses.

Thousands of schoolchildren—up to one half of the 19,000 student sample—were randomly chosen to be experimental "controls" in the eyesight experiment (it is difficult to determine from the study precisely the number of students selected for the control group). Students randomly selected to serve as "controls" were not supplied with corrective eyeglasses nor any other eye care, however sight defective and prepared they were to benefit from the best practice treatment, if given.

Students in the control group were recruited and followed and tracked, same as the "treatment" group, but under no circumstance were the "controls" to be given corrective eyeglasses (Glewwe, Park, and Zhao 2012, p. 8). "The lack of rigorous studies on the impact of providing eyeglasses to students with visual impairments in developing countries led to the implementation of the Gansu Vision Intervention Project in 2004 in Gansu Province in northwest China."

Why were thousands of poor Chinese schoolchildren with defective eyesight recruited and followed for a full calendar year (p. 8) and yet not given corrective eyeglasses to wear at school? Because the experimental economists (by subfield name the "development" economists, exactly, technically speaking) wished to "test" against nil placebo whether wearing corrective eyeglasses might improve children's educational performance.

No one is sure why economists and their sponsors at the World Bank, National Institutes of Health, and Spencer Foundation had any doubt about the null hypothesis, which is blindingly obvious (Glewee, Park, and Zhao, acknowledgements). Helen Keller said, "It is a terrible thing to see and have no vision." We wonder why the sample size had to grow to 19,000 students before the trial was considered valid. Conduct the eyesight/school productivity experiment on yourself, in your own classroom (n=1). Take off your glasses and try to read a smattering of Malthus or mathematics scribbled over there on the blackboard. If you are far sighted instead of near sighted, there is a difference. But the large negative effect is still there; sight defective professors need glasses to improve their scholastic work, and so do sight defective children. If you are still in doubt after n=1, ask your teenage baby sitter: "Does wearing glasses help?" (we mean for school, not vanity). Regardless, you do not need 19,000 schoolchildren from any nation to reject the null hypothesis of "no help".

To test the hypothesis that eyeglasses help at school, the authors of the Eyesight Experiment argued along with most other experimental economists that there is a need for a no treatment control group, meaning that some of the kids can't have the glasses.

Some data were dropped for fear of spoiling the experimental design. "Unfortunately," the authors write, "in a few cases students in control townships were given eyeglasses because,

[5] None of the studies in our survey of the American Economic Review, 2000-2009, considered equipoise and "no treatment" controls from the perspective of actual study participants—an odd fact about a discipline that is otherwise devoted to maximum welfare, freedom of choice, and methodological individualism. Clinical researchers have paid far more attention to the role of individual values and preferences: see, for example, Alderson (1996) and the Bayesian approaches of Kadane (1986) and Lilford and Jackson (1995).

after providing the eyeglasses in the treatment townships, the remaining funds were used to buy eyeglasses for students with poor vision in the paired control township. This occurred in two control townships in Tianzhu and three control townships in Yongdeng" (Glewwe, Park, and Zhao 2012, p. 8). In remaining townships the "randomization was done according to the plan" (Glewwe et al. p. 8), the authors of the China study state. But one ought to question a practice which elevates abstract method over ethics and the chance to really help.

Substituting syphilis for short-sightedness, we are reminded of the Tuskegee Syphilis Experiment, 1931-1972, wherein doctors employed by the U.S. Department of Public Health were, they believed, advancing science by deliberately *not* treating hundreds of syphilitic African American men in Macon County, Alabama (Jones 1992; Gray 1998). "The doctors [from the U.S. Department of Public Health] were sure that untreated syphilis was a deadly problem and that treatment [penicillin] was efficacious, that they wanted to prove it beyond question by control group comparisons and autopsies that would rule out any other possible explanation" (Gray 1998, p. 94). Short-sightedness is not deadly but it is highly treatable.

In medical ethics such "no treatment" studies violate "the personal care principle" (Royall, 1991) – an oath accepted by physicians to provide best available treatment. "No treatment" control groups are thus ethically suspect. In economic ethics "no treatment" disturbs, for example, the general rules of conduct laid down by Adam Smith's (1791 [2009], p. 280) impartial spectator, who cannot accept the utilitarian reply that the deliberate sacrifice of today's children is for the betterment of tomorrow's. A transcendental theorist of justice, rights, and duties—such as the followers of John Rawls and John Rawls himself—would agree with Smith. And so, too, would Amartya Sen (2006, p. 217), who judges experiments much like Smith did, not from behind a veil of ignorance but empirically using the "comparative" (p. 217) method:

> Investigation of different ways of advancing justice in a society . . . or of reducing manifest injustices that may exist, demands comparative judgments about justice, for which the [transcendental Rawlsian] identification of fully just social arrangements is neither necessary nor sufficient. To illustrate the contrast involved, it may well turn out that in a comparative perspective, the introduction of social policies that abolish slavery, or eliminate widespread hunger, or remove rampant illiteracy, can be shown to yield an advancement of justice.[6]

In truth, the development economists could have used their millions of dollars of funds to simply purchase eyeglasses for each of the thousands of untreated children discovered by them.[7] This would not serve a purpose for abstract research but it would serve a higher purpose, Sen would agree, by helping poor and needy schoolchildren with best practice (and affordable) treatment. To illustrate, at the time of the eyeglass experiment in China, one pair of correctives

[6] For Rawls (1971, p. 83) inequality is permissible if and only if the inequality can be shown to benefit the least advantaged persons. The treatment/no treatment inequality does not.

[7] We were not able to determine the cost of the eyesight experiment. But randomized trials are rarely low cost. Johnston et al. (2006, p. 1319) report in *The Lancet* on 28 large scale trials costing $335 million or $12 million for the average trial. Only six of the trials (21%) resulted in measurable improvements for trial participants and only 4 of the 28 trials (14%) resulted in cost savings to society. By 2004 the U.S. National Institutes of Health was investing $3 billion annually (p. 1319) in these and other large scale clinical trials.

cost on average the equivalent of about $15.00 U.S. (nominal 2012 dollars; Glewwe, Park, and Zhao, p. 28). Lower prices for glasses were then available in China, for as low as $2.00 U.S. per pair per student. So with an expenditure equal to or less than the cost of the experiment the World Bank, NIH, and other grant money could have been used a lot more efficiently and a lot more justly by empowering thousands of additional students. (Needless to say, the authors conclude from their study that corrective eyeglasses give a significant boost to school performance.)

The authors decided not to work at all with children and schools located in remote areas of rural China, meaning that the poorest of the poor did not get eyeglasses from the experiment. "Primary schools with less than 100 students were excluded from the project to avoid high travel costs to a few very remote schools. Students in such schools are only 6% of primary students in the two counties. . . This leaves six pairs of townships in Yongdeng and six pairs (plus the poorest township, the one randomly assigned to be treated) in Tianzhu for which the randomization was done according to the plan."

II. The Mostly Random Rise of Randomized Trials in Economics

"I am a huge fan of randomized trials," Hal Varian (2011) told *The Economist.* "[T]hey are well worth doing since they are the gold standard for causal inference," he asserts without proof. Economists working for Varian, the chief economist of Google, "ran about 6,000" randomized trials at Google in the year 2010 alone. A leading *randomista*, the John Bates Clark Award-winning economist, Esther Duflo, has told *The New Yorker* that field experimentalists [such as her and her colleagues at MIT] have borrowed from medicine a "'very robust and very simple tool' . . . they subject social policy ideas to randomized control trials, as one would use in testing a drug. 'This approach,' Duflo claims, 'filters out statistical noise; it connects cause and effect'" (quoted in Parker 2010, pp. 79-80). List (2008) agrees with the drug-testing analogy, and so do many others in the current generation.

Levitt and List (2009) go further and assert that the introduction of randomized treatments and controls—of completely randomized blocks—laid the "foundation" for good experimental design (Levitt and List (2009, p. 3). Artificial randomization of treatments and controls is, they claim, the only "valid" justification for use of Student's test of statistical significance.[8]

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).[9] Levitt and List consider this quote from Fisher (1935) the foundation of experimental method:

[8] Levitt and List, p. 3; contrast Ziliak (2014, 2011b, 2008), 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).

[9] The advocates of RCTs in economics appear to be innocent of the real sources of randomized trials in medicine, the field they claim to imitate. For example, in addition to their neglect of Student, none of the new generation makes any mention of A. Bradford Hill and his immense influence on the use of RCTs in medicine. See the recent symposium introduced by Iain Chalmers (2003). Chalmers (pp. 922-924) neglects Student, too, repeating the popular but incorrect Fisherian history of randomized trials.

> 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.[10]

"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" (Levitt and List, p. 3).

To Fisher and today's *randomistas* blocks (or strata) have the same probability of being "treated alike" only when treatments and controls are randomly assigned to the experimental unit. Yet Duflo, Banerjee, Karlan, List, Levitt and others—following Cochrane (1976), Rubin (1990), and Street (1990)—do not explain why Fisher is to be believed.

The enthusiasm for randomized trials is not limited to academics. The World Bank asserts in a research guidebook that randomized trials are the most "rigorous" type of assessment (World Bank 2004, p. 23). The United Nations Food and Agriculture Organization published back in 1999 a 234 page long Statistical Manual for Forestry Research, "Design and Analysis of Experiments" (Jayaraman, 1999). On randomization the FAO sounded the usual bell:

> Assigning the treatments or factors to be tested to the experimental units according to definite laws or probability is technically known as randomization. It is the randomization in its strict technical sense that guarantees the elimination of systematic error. It further ensures that whatever error component that still persists in the observations is purely random in nature. This provides a basis for making a valid estimate of random fluctuations which is so essential in testing of significance of genuine differences. . . Through randomization, every experimental unit will have the same chance of receiving any treatment.

III. Three Big Losses Caused By Randomization

But the FAO and World Bank are, like many academics, not as aware as they might be of several big losses caused by randomization. Randomization in the design of an experiment is normally achieved by using a random number generator to allocate treatments and controls to experimental units. For example, a barley farmer may wish to test the hypothesis that, other things equal, crop yield is importantly higher when crops are fertilized—the unfertilized crops serving as controls. There are at least three reasons why the rational statistician would rather balance and stratify treatments and controls rather than completely randomize over the experimental unit.

1. *Randomization leads to Simpson's Paradox, reversing the sign of causal effects*

Let T be a treatment in an experiment to be tested against control P, the company or industry standard, a sugar pill, or—as in many field experiments in economics—no treatment at all.

[10] Fisher (1935, p. 26).

Suppose you run an experiment on n=80 adults, randomly assigning the treatment *T* to some of the randomly chosen adults and control *P*—which, as in the Chinese eyeglass experiment, economists often equate with "no treatment at all"—to the others in the sample, whom we will call the control group.[11] Here are the aggregate results:

Results for all Participants

		S	F	Total	Success Rate
Treatment	*T*	40	40	80	50%
	P	32	48	80	40%

where S = treatment success (for example, an unemployed person gets and keeps a job for a certain number of months while receiving either all treatment *T* or all control *P*), and where F = treatment failure (for example, the unemployed did not get a job, the sight defective student did not scholastically improve).

The aggregate results appear to show that treatment *T* works much better than the control *P*—a 50% success rate versus a 40% success rate. But the aggregate results assume that the participants are homogeneous in all important fixed factors and covariates (income, gender, race, skill level, education, et cetera). When the units of the experiment are made more homogeneous, by disaggregating the data into strata by gender, for example, the results are as follows:

Results for Stratum 1 (Men)

		S	F	Total	Success Rate
Treatment	*T*	36	24	60	60%
	P	14	6	20	70%

Results for Stratum 2 (Women)

		S	F	Total	Success Rate
Treatment	*T*	4	16	20	20%
	P	18	42	60	30%

Note: Results for stratum 2 = Aggregate results – Stratum 1

But now, after controlling for the heterogeneity caused by gender difference, the balance of evidence has shifted, reversing the result of the experiment. The treatment *T* is not the best, most successful treatment after all – not when judged from the perspective of just two relevant strata, male and female. For both men and women, the disaggregated data show that the control *P* dominates the treatment *T* (that is, the percentage success rate is higher in the control group)

[11] The data were supplied by Lindley (1991, pp. 47-48) but Lindley does not mention that Simpson's Paradox can be caused by artificial randomization of treatments and controls.

making it clear that treatment *T* is effective for the whole sample but harmful for both men and women.

Something is wrong. It is called Simpson's Paradox, a common defect of statistical studies and especially of randomized controlled trials. In summary, if the experiment is not prudently stratified to eliminate heterogeneity bias, the randomized trial can mislead investigators. For example, in the Chinese eyesight experiment complete randomization suggests that 73% of the boys accepted glasses when offered but only 66% of the girls accepted (Glewwe et al., p. 30). Further analysis of the sample shows, however, that boys were offered glasses 8.5% more frequently than girls. Thus a stratified sample might serve to reduce or eliminate the apparent difference in propensity to accept eyeglasses. Regardless, the authors uncovered a much larger strata of difference, spoiling the external validity of their study: the social position of parents. Turns out that children of schoolteachers and village cadres were much less likely than average to accept the eyeglasses when offered:

> children in households headed by a schoolteacher or a village cadre were less likely to accept glasses . . . These effects are very large, with schoolteachers' children 22.4 percentage points less likely, [and] village cadres' children 35.2 percentage points less likely, to accept them. Perhaps these local authority figures decline program benefits to avoid being perceived as manipulating the program for personal benefit. Alternatively, it would be strange, and ironic [the authors admit], if these authority figures had more doubts about the merits of eyeglasses.

2. *Randomization raises the probability of imbalance, biasing estimation and tests of significance, and possibly reversing the impact of treatments and controls*

A second major flaw in randomized trials is caused by a related design failure – the failure to control for allocation imbalance and systematic error. Imbalance can occur for at least two reasons: first, the treatments might be given disproportionately to one strata and not to the others, as in the example above, where the treatment *T* was given disproportionately to men (60 men, 20 women). Second, imbalance can occur when there are unobserved omitted variables in the system that are large with respect to the output of interest, and yet are systematic and not able to be artificially randomized.

Consider a classic example from agricultural economics and experimental statistics: designing a yield trial to compare the estimated yield per acre of barley variety A (the treatment) versus barley variety B (the control), when there is differential soil fertility cutting across the farm plot, the experimental unit. We need some way or other to allocate seeds A and B to rows and columns of the plot (or plots, plural – ideally speaking, in a small sample of independent and repeated experiments: Ziliak (2014, 2011b). Suppose each plot is divided into blocks or sub-blocks. For simplicity, imagine that each block gets a unique "treatment", A or B, determined by random coin flip, head for A and B for tail. The problem with random assignment of treatment and control to natural soil is that by a series of chance coin flips, the recommended allocation of As and Bs could turn out highly imbalanced with regard to a major variable of crop growth, soil fertility:

A A A A A A B B
A A A A A A B A (i)

. . . . etc.

→ Direction of increase in soil fertility (higher yielding soil)

And a second series of random flips could produce an imbalanced allocation of this sort:

B B B B B A A B
B B B B B A A A (ii)
. . . etc.

→ Direction of increase in soil fertility (higher yielding soil)

How precise are the estimated differences in average yields, A-B, or B-A, if fertility on the left side of the field is systematically lower than fertility on the right? Layouts such as (i) and (ii)—though randomly chosen by the field experimentalist—produce biased mean squared errors and parameter estimates with respect to this major source of fluctuation—differential soil fertility (cf. van Es, et al., 2007; Meyers, et al., 2011). Likewise in the Chinese eyesight experiment the randomized sample is found to be highly imbalanced with respect to a large non-random effect: the differential behavior of the schoolchildren's parents, which varies greatly, the authors found, by occupational level and social status. Thus, as Heckman would note, the eyeglass experiment suffers also from selection bias.

Student proved the point first, we've noted, in Guinnessometrics and agronomy. In example (i) the As are bunched up and growing in the very worst soil; thus the yield of the Bs will be artificially high, and the real treatment difference, A-B, will be undetermined. Student 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 2 error and errors from fixed effects, such as non-random soil heterogeneity.[12]

As Student showed (1938, 1923, 1911) examples (i) and (ii) suggest that whenever there is a systematically variant fertility slope (or other temporal or secular source of local and fixed effect, as there so easily could have been in the Chinese experiment) 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 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 Student's solution became perfectly balanced. The ABBA layout is:

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

[12] Student (1938, pp. 364-372).

One virtue of the ABBA *non*-randomized design is that it minimizes bias caused by differential soil fertility. Student found that the "principle of maximum contiguity" (as he called it) takes full advantage of correlation of variables, by twinning in the manner of Noah's Ark, and holding other things equal. Given the built-in symmetry of ABBA, no matter what the trajectory or magnitude of differential fertility gradients, A's and B's are equally likely to be grown on good and bad soil. Random throws of seed do not have this virtue, biasing mean yield differences, A-B.

Yet ABBA brings additional statistical and economic advantages, too.[13] On the supply side, with ABBA the ease and cost of sowing and harvesting and calculating basic statistics on yield is plot-wise and block-wise reduced. Compare the rows and columns of ABBA with the random rows and columns in (i) and (ii) above and it's easy to appreciate Student's sensitivity to supply side economic conditions.

With ABBA there is no need for chaotic tractor driving while planting seed in blocks randomly dispersed; and thus with ABBA there is a lot less measurement error and loss of material at harvest and counting time (see Beaven, 1947 for 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."[14] 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 refused to admit the economic and statistical advantages of Student's ABBA and other balanced designs (see, for example, Fisher and Yates (1938), which fails to mention Student's methods). In 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:[15]

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

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

[15] Student (1938, p. 366).

> 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.

As far back as 1926, Student (1926, p. 126) explained the flaws of randomization to agricultural economists:

> Generally speaking, . . . the population of large-scale yields with which we are concerned is a population of "differences", i.e., some such question as the following is asked: "By how much may we expect the yield of variety B to exceed that of variety A if they were sown alternatively on the same soil in the same season?" . . . That being so, it is clear that the observed differences will not represent the true differences even in the same sample plots as two crops cannot occupy the same place at the same time. Observed differences will miss the mark not only because the experimental soil and the weather experienced by *the experiment may not be random samples of the soil and weather to be explored, but also because the actual plots laid out for the two varieties will usually differ in fertility. This is one of the largest sources of errors in field experiments"* (Student 1926, p. 126; emphasis added).

Despite these warnings, today's *randomistas* follow R.A. Fisher. They have not yet perceived that Student's balanced designs and small series of independently repeated experiments continue to dominate random.

In general, Student's Imbalance and Simpson's Paradox can spoil results from any randomization scheme, including instrumental variables with non-experimental data, which is said to solve the problem of omitted variable bias. Angrist and Krueger (2001, p. 72) are, for example, mistaken when they assert:

> One solution to the omitted variables problem is to assign the variable of interest randomly. For example, social experiments are sometimes used to assign people to a job training program or to a control group. Random assignment assures that participation in the program (among those in the assignment pool) is not correlated with omitted personal or social factors . . . Instrumental variables offer a potential solution in these situations.[16]

[16] See also: Angrist and Pischke (2009, pp. 15-17).

3. *Randomized controls fail at the margin of economics - and ethics, too*

Thirdly, if randomization of treatments and controls has economic and/or ethical advantage over balanced or other systematic designs, then pure randomization would win at the margin.

Decisions based on randomized allocations would be, other things equal, more valuable at the margin than would alternative, deliberately made decisions using systematic or what is known as balanced designs of experiments. Ethically speaking, if randomized controls are preferred, it would be easy to prove that withholding best practice treatment raises the well-being and life chances of participants assigned to the untreated control group.

That is not always the case. Consider making a decision at the margin of the Chinese eyeglass experiment. Suppose you, the development economist, have $15.00 to spend on each study participant, and you have now to choose one of two options to spend it.

Option 1: You supply a pair of prescription eyeglasses (costing $15.00) to all sight defective children you encounter in your study, given that they do not have eyeglasses of their own. You follow them in school and note the improvement.

Option 2: You flip a fair coin each time you meet a sight defective child in the study. If the coin turns up head, give the child a pair of eyeglasses; if tail, do not supply the glasses; instead spend the money tracking and reporting on untreated students.

Does Option 2 – the option and method frequently employed by today's randomizers - feel ethically correct? Or does the ethical mandate of the personal care principle (or, alternatively, of the impartial spectator) dominate "no treatment" at the margin? Our "feelings" for the untreated children or other subjects of economic and medical trials may not be the decisive factor. But they are certainly one of the deciding factors the ethical economist must consider. In Part II, Section III of *The Theory of Moral Sentiments*, "On Merit and Demerit; or, Of the Objects of Reward and Punishment," Adam Smith (1791 [2009]) observes:

> Whatever praise or blame can be due to any action, must belong either, first, to the intention or affection of the heart, from which it proceeds; or, secondly, to the external action or movement of the body, which this affection gives occasion to; or, lastly, to the good or bad consequences, which actually, and in fact, proceed from it. These three different things constitute the whole nature and circumstances of the action, and must be the foundation of whatever quality can belong to it.

What is missing now is a sense of balance in both ethics and statistics. A balanced ethics would give, for example, best practice treatment to the control group, while trying out on willing others a novel and promising treatment about which not much is known. Casey Mulligan (2014) agrees that if research workers insist on using a no treatment option, the cost of that decision should be shared by them, with skin in the game, such as by paying more cash money to people who are willing to be randomly chosen for their "no treatment" control group. Currently, most randomized trials fail to meet a basic postulate of welfare economics, Pareto efficiency (see, for example, Johnston, et al. [2006]). The defensive answer which is often heard in reply is "Policy makers have to be convinced of the initiative. How can we convince policymakers that investment in [Project X] is worthwhile?"[17]

III. The So Called Randomization Principle Is Not a Principle

Other economic statisticians have gone far beyond mere randomization and randomized controlled trials (Kadane, 1986). For example, statisticians have long known that stratification, or blocking, is a first necessary step to improving the precision and efficiency of a study based on pure randomization. Student (1911) used blocking and stratification before the synonymous words existed in the statisticians' vocabulary. As W. Edwards Deming (1978, p. 879), an admirer of Student, noted, "Stratification is equivalent to blocking in the design of an experiment."

Box, Hunter, and Hunter (2005, p. 92) explain that "A block is a portion of the experimental material (the two shoes of one boy, two seeds in the same pot) that is expected to be more homogenous than the aggregate (the shoes of all the boys, all the seeds not in the same pot). By confining comparisons to those within blocks (boys, girls), greater precision is usually obtained because the differences associated between the blocks are eliminated." Blocks are strata.

Deming (1978), who before turning to manufacturing did a long stint at the U.S. Department of Agriculture, agreed with Student's larger point: 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 or when the cost of being wrong is negligible. Said Deming (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.[18]

[17] Martin Ravallion (this volume), a former director of research at The World Bank, has been a critic of randomized trials. He told the authors that he would defend the Chinese eyesight experiment and other experiments like it on, he said, "consequentialist" grounds (Ravallion, 2014). "Actually, I think the consequentialist argument is key," he argues. But the ends do not justify the means for at least two reasons. First, the means of withholding best treatment from impoverished people today, to possibly help unknown others in the future, is not justified by any ethic save vulgar utilitarianism, which both Ravallion and we reject. Second, the ultimate ends of the experiment are, Hayek and others would note, evolving and unknown in the developing Chinese economy.

[18] Deming (1978, p. 879). Deming said he learned the technique from Neyman (1934). In the seminal article Neyman proves mathematically and empirically the statistical and economic advantages of stratified sampling over random sampling (Neyman 1934, pp. 579-585). Neyman credits the idea of "purposive selection" to earlier writers, such as Bowley and Gini and Galvani.

Likewise in his book, *Planning of Experiments,* David Cox (1958) recommends "completely randomized arrangement . . . [only] in experiments in which no reasonable grouping into blocks suggests itself"—that is, when ignorance prevails, or priors are flat.

Normally speaking ignorance does not prevail, and real economic and statistical gains can be found by stratifying. Deming (1978) and Tippett (1952) simplified Student's (1911, 1923) proof that stratification (blocking) can reduce sample size requirements by 40% or more, holding variance constant.[19] And 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."

According to Levitt and List (2009, p. 4) "Fisher and McKenzie (1923)" is the second classic article to use randomization in the design of a field experiment. This is a remarkable achievement given that randomization does not appear even once—randomization is neither used nor mentioned—in the article by Fisher and Mackenzie. "Fisher's fundamental contributions were showcased in agricultural field experiments. In his 1923 work with McKenzie, Fisher introduced . . . randomization" (Fisher and McKenzie, 1923)," Levitt and List write. 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.

As late as 1923—the same year that Student was comparing random with balanced designs on farm fields across England and Ireland—Fisher had not given much thought to the statistical design of experiments. Cochrane (1989, p. 18) notes that: "Fisher does not comment [in Fisher and Mackenzie (1923)] on the absence of randomization or on the chessboard design. Apparently in 1923 he [that is, Fisher] 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 (Fig. 1) [by Fisher and Mackenzie] illustrates [Yates said] how little attention was given to matters of layout at that time.[20] It is indeed difficult to see how the arrangement of the varietal plots [designed by Fisher and Mackenzie] was arrived at.

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 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

[19] Deming (1978, p. 880-881), Tippett (1958, p. 356). In a Riesling vine-and-wine experiment, Meyers, Sacks, van Es, and Vanden Heuvel (2011) used blocking, balancing, and repetition (at n=3 vineyards) to reduce sample size requirements by up to 60%.

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

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). *Randomistas* want their readers to believe otherwise.

Yates continued (Yates 1964, pp. 311-312): "The principle of randomisation was first expounded in [Fisher's textbook] *Statistical Methods for Research Workers* [1925]"—not in Fisher and Mackenize (1923).[21] In other words, an article that Levitt and List claim for the canon in the history of randomization neither mentions nor uses randomization. Besides, randomization was proven to be inferior to balanced designs and anyway it is not even a principle. "I don't agree with your controlled randomness," Student told Fisher in a letter of October 1924 (Gosset, 1962). "You would want a large lunatic asylum for the operators who are apt to make mistakes enough even at present," Student said of Fisher's so-called 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 make a note.

IV. Survey of Randomization, Significance, and Validity in Economics and Medicine

To better understand randomization in practice, we designed a survey of randomization, statistical significance, and validity as used in randomized controlled trials in economics and medicine. We put 25 questions to published articles, eliciting Yes, No, and Not Applicable answers, where "Yes" indicates that the study corrected for or at least acknowledged the potential costs of randomization bias and the cult of statistical significance. Several of the survey questions are directly comparable with surveys of significance testing in general, as conducted and reported by Ziliak and McCloskey (2008; McCloskey and Ziliak 1996). The survey questions are drawn from the past century of best practice experimental statistics, from the works of William S. Gosset aka "Student" to James J. Heckman, not including Ronald A. Fisher.[22] Given our ethical concerns about randomization, we also inquire whether each study established a no treatment control group when in fact known treatments were available at the time of the study (Question 1).

Tables 1, 2, and 3 present survey results for the last decade of publishing randomization studies in the *American Economic Review*, January 2000-December 2009. We compare the AER results with the population of full-length articles using randomization techniques and published in the *New England Journal of Medicine*, 2000-2002.

Here are the results from our survey:

[21] The case for random arrangement is made by Fisher (1925, Sections 48 and 49, "Technique of Plot Experimentation" and "The Latin Square," pp. 229-237); for the full story, see Ziliak and McCloskey (2008, chps 20-22).

[22] Excluding Fisher's philosophy from the survey design is natural. The reason is not because he was not a major figure in the development of statistics. Clearly he was. It's because Fisher's philosophy and rhetoric was, we have noted elsewhere (Ziliak, 2014, 2011b, 2010a; Ziliak and McCloskey 2008, chps. 20-23), a major cause of today's randomization and significance school.

*All of the articles published in the AER (100%) fail to provide any information at all as to the balance or lack of balance of covariates and treatments in their experiment (Table 2, Questions 4a and 4b). That is a lot of occasions for effect reversal from Simpson's Paradox, from Student's Imbalance, and from other systematic errors of judgment under conditions of uncertainty;

*A mere 14% of the AER articles stratified the sample, producing, as Student and Deming and Cox noted long ago, a lack of confidence in randomly generated results (Table 2, Question 3); practice was better in the NEJM (40% stratified) but that still leaves 6 of every 10 medical studies at risk of misleading conclusions from Simpson's Paradox and Student's Imbalance;

* None of the AER articles (0%) and less than 1% of the NEJM articles focused on substantive "size matters/how much" questions (Table 1 and Table 2, Questions 16-21). *Randomistas* base experimental conclusions exclusively on tests of statistical significance which reach or fail to reach the arbitrary 5 or 10% level of significance, ignoring the "oomph" (Ziliak and McCloskey, 2008);

*11% of the AER articles and only 1% of the NEJM articles replicated a previous experiment, suggesting that more attention might be paid to establishing the external validity of randomized trials (Table 1 and Table 2, Question 5);

*None of the NEJM articles (0%) and only 5% of the AER articles did any graphing or other diagnostic tests on the distribution of error terms and of correlations to determine by how much their data depart, if they do, from the normal assumptions of internal validity (Table 1 and Table 2, Question 7);

*None of the NEJM articles (0%) and only 1% of the AER articles considered other sources of evidence before concluding that their statistically significant finding from one randomized set of data is "valid" (Table 1 and Table 2, Question 23); this finding confirms what the randomizing experimentalists themselves, from Banerjee to List, have claimed: they do not value evidence produced by other types of studies—econometric, historical, and other (contrast Harrison [2011], who does).

*Only 5% of the AER articles explain the choice and likely implications of their sample size selection, and thus of the power of their tests (Table 2, Question 15 and Question 20). The Guinness Brewery, and Student himself, would be puzzled: don't we need to know the economically most efficient way of learning?

*Only 1% of NEJM articles include data on costs and benefits related to the experiment such that one might estimate the net benefit of the authors' favored hypotheses (Table 1, Question 13); authors of the AER performed better (though not exemplary) in this regard, with 44 of 80 or little more than half of the economics articles providing basic information about cost and benefit (Table 2, Question 13).

*More than two-thirds (69%) of the AER studies were sponsored by a government, a business, or a private grant-making agency but none of them (0%) made a conflict of interest statement (Table 2, Questions 24 and 25).[23]

We see these and other problems played out in the works of our colleagues. On the question of validity, for example, consider the article by Salazar-Lindo, Sebastian-Ponce, et al. (2000), published in the NEJM. The article is the end result of a four-year study to test the benefits of a drug, racecadotril, for treating acute watery diarrhea in children. The authors studied exclusively male infants (aged 3-35 months) with diarrhea but they extrapolate their results onto adults, a leap of faith. They rely on controlled tests, though we applaud them for using the next best alternative treatment for diarrhea rather than "no treatment," as economists are prone to do with poor people in studies from eyeglasses to mosquito nets, surprising in a field emphasizing the next best alternative as the definition of cost.

Salazar-Lindo, et al. compares favorably with a study in the United Kingdom of antenatal steroids and infant mortality, first published back in 1972. "The outcome of interest was infant mortality due to complications of immaturity [in the womb]" (Howick 2011, p. 18). Pointless debates about the "significance" and "insignificance" from a number of independently randomized controlled trials dragged on until 1995, delaying marketing of the steroid for pregnant mothers and causing many preventable deaths. In a systematic review of the combined studies (4 in the final analysis) Patricia Crowley discovered that 1000 of the premature babies (and their mothers) received the antenatal steroid and about the same number, 1000, received a placebo. Of the babies who received the steroid, "70 died, while 130 [about twice as many] died in the placebo group" (Howick, p. 19).

In Fehr and Goette (AER, 2007) the authors seek to find the effect of transitory wage changes on workers' labor supply by studying two Swiss bicycle messenger services and artificially manipulating the commission paid to messengers. But as usual in both the AER and NEJM, there are no diagnostic checks for validity, whether internal or external. And there is no evidence of balance and stratification. Fehr and Goette studied bicycle messengers at two companies in Switzerland. Should readers assume their merely statistically "significant" findings are representative of the whole Swiss labor force? What, if anything, can we learn about American or Belgian or Indian bicycle messenger services?

Of the 80 articles using randomization techniques in the AER over the past decade, 21 of 80 or 26% are controlled field experiments wherein subjects are randomly selected by economists and offered one of two or more treatments, including a "no treatment" or nil placebo control group in one half (52%) of the real world trials (Table 2, Question 1a). By "no treatment" we mean those studies in which a preferred treatment (a low cost HIV test, a food voucher, a pair of eyeglasses) is experimentally withheld from a control group. For example, Jensen and Miller (2008) withheld food vouchers from poor Chinese farmers, and Thornton (2008) withheld HIV test vouchers from randomly selected African villagers living in Malawi.

The practice of giving no treatment at all appears to be more pronounced in the field journals of development and applied economics. For example, in 2013 in a sister journal, *American Economic Journal: Applied Economics*, edited by Esther Duflo, 7 of the 9 (that is

[23] On January 5th, 2012 the American Economic Association adopted "extensions to [seven added] principles for author disclosure of conflict of interest," including disclosure of bank and corporate sponsorship, if any. See Epstein and Carrick-Hagenbarth (this handbook).

77%) of the randomized controlled trials published offered no treatment to the control group. Again, we found some pushback against “no treatment,” and again, the pushback came not from the economists but from the moral conscience of local teachers and staff cooperating with the economists. Fairlie and Robinson (2013, p. 215), for example, is a randomized controlled experiment designed to see if having a computer at home affects student academic success: "In discussing the logistics of the study with school officials, school principals expressed concern about the fairness of giving computers to a subset of eligible children. For this reason, we decided to give out computers to all eligible students. Treatment students received computers immediately, while control students had to wait until the end of the school year. Our main outcomes are all measured at the end of the school year, before the control students received their computers". The practice of withholding best practice treatment from the control group is in drug and medical experiments equally pronounced. Of the 234 randomized trials published in the New England Journal of Medicine, nearly half (48%) offered participants in the control group a placebo or no treatment at all (Table 1, Question 1). None of the studies we examined discuss or justify the practice of withholding best treatment from the control group.

In short, it would seem that a large majority of randomized field trials in economics and medicine are economically, experimentally, and ethically speaking out on a limb.

{ NOTE TO EDITOR: Insert Tables 1,2, 3 about here }

TABLE 1: RESULTS OF NEW ENGLAND JOURNAL OF MEDICINE SURVEY OF RANDOMIZATION, SIGNIFICANCE, AND VALIDITY (2000-2002), RANKED BY PERCENT YES

Survey Questions	Number of Applicable Articles	Percent Yes
Does the study…		
7. Do diagnostic tests to assess “internal validity”?	234	0
7a. If the study does diagnostic tests to asses "internal validity" does it: do nothing (0), plots residuals for test of normally distributed error terms (1), or tests for independence of error terms (2)?	234	0
16. Acknowledge that statistical significance (or the lack of it) is not decisive?	234	0
18. Specify a loss function or otherwise stipulate the meaning of effect sizes, such that the reader can assess	234	0

the how “large is large” and how “small is small” questions?		
19. If it does not specify a loss function, does the study discuss otherwise the implications of being wrong about the preferred hypothesis?	234	0
21. Seek mainly to demonstrate substantive significance?	234	0
23. Consider the validity of other types of evidence, that is, evidence not based upon formal experiment and artificial randomization?	234	0
5. Replicate a previous RCT or other types of empirical study?	234	1
9. Discuss its design relative to previous studies of the same?	234	1
13. Include cost and benefit data, such that one can determine the real net benefit of the experiment- adverse effects included?	234	1
17. Recognize that statistically insignificant results might have large and important substantive effects?	234	1
14. Include cost and benefit data such that the reader can estimate the net benefit of accepting the preferred hypothesis?	234	6
12. Compare results (such as effect sizes, magnitudes of coefficients, etc.) with previous studies?	234	14
2. Strive to test and estimate multiple treatments?	234	26
25. Provide conflict of interest information?	234	39
3. Design balance covariates (confounding variables) prior to experimentation? Or, if the study is an observational study, does it stratify the sample prior to data collection?	234	40
15. Explain the choice and implications of sample size?	234	48
20. Discuss the power of the test?	234	48
22. Mimic conventional best practice, in both experimentation and implementation, in the industry for the “treatment” in question?	234	51
1. Eschew the use of placebo control and no treatment?	234	52
6. Test hypotheses (or treatments and controls) at multiple places?	234	82
8. Does the study test the favored model on sub-populations or strata and/or test for robustness of estimates across different periods of time?	234	83

4a. Does the study discus questions of balance for sample strata and covariates?	234	84
4. Eschew balancing (and/or stratifying) after experimentation?	234	88
24. Is the study sponsored by a government, firm, or other non-academic entity?	234	93
4b. Does the study give estimated magnitudes of balance (or imbalance) in table, chart, or other form?	234	93
10. Defend the design of the experiment; for example, does it explain why it's using artificial randomization? Or, does the study explain eligibility criteria for participants' inclusion?	234	99
11. Provide sufficient details to allow replication- including details on how and when treatments and controls were actually administered?	234	100

Source: All full-length articles that use randomization published in the *New England Journal of Medicine*, January 2000 – December 2002.

Note: "Percent Yes" is the total number of Yes responses divided by the relevant number of articles.

TABLE 2: RESULTS OF AMERICAN ECONOMIC REVIEW SURVEY OF RANDOMIZATION, SIGNIFICANCE, AND VALIDITY (2000-2009), RANKED BY PERCENT YES

Survey Questions	Number of Applicable Articles	Percent Yes
Does the study…		
4a. Does the study discuss questions of balance for sample strata and covariates?	80	0
4b. Does the study give estimated magnitudes of balance (or imbalance) in table, chart, or other form?	80	0
16. Acknowledge that statistical significance (or the lack of it) is not decisive?	80	0
17. Recognize that statistically insignificant results might have large and important substantive effects?	80	0

18. Specify a loss function or otherwise stipulate the meaning of effect sizes, such that the reader can assess the how “large is large” and how “small is small” questions?	80	0
19. If it does not specify a loss function, does the study discuss otherwise the implications of being wrong about the preferred hypothesis?	80	0
21. Seek mainly to demonstrate substantive significance?	80	0
25. Provide conflict of interest information?	80	0
20. Discuss the power of the test?	80	1
23. Consider the validity of other types of evidence, that is, evidence not based upon formal experiment and artificial randomization?	80	1
7. Do diagnostic tests to assess “internal validity”?	80	5
7a. If the study does diagnostic tests to asses "internal validity" does it: do nothing (0), plots residuals for test of normally distributed error terms (1), or tests for independence of error terms (2)?	80	5
15. Explain the choice and implications of sample size?	80	5
5. Replicate a previous RCT or other types of empirical study?	80	11
3. Design balance covariates (confounding variables) prior to experimentation? Or, if the study is an observational study, does it stratify the sample prior to data collection?	80	14
6. Test hypotheses (or treatments and controls) at multiple places?	80	20
22. Mimic conventional best practice, in both experimentation and implementation, in the industry for the “treatment” in question?	60	20
10. Defend the design of the experiment; for example, does it explain why it’s using artificial randomization? Or, does the study explain eligibility criteria for participants’ inclusion?	80	23
12. Compare results (such as effect sizes, magnitudes of coefficients, etc.) with previous studies?	80	26
14. Include cost and benefit data such that the reader can estimate the net benefit of accepting the preferred hypothesis?	80	29

8. Does the study test the favored model on sub-populations or strata and/or test for robustness of estimates across different periods of time?	80	36
13. Include cost and benefit data, such that one can determine the real net benefit of the experiment- adverse effects included?	80	44
2. Strive to test and estimate multiple treatments?	80	46
1a. Eschew the use of placebo control and no treatment? (All Field Experiments, not Laboratory)	21	48
9. Discuss its design relative to previous studies of the same?	80	60
24. Is the study sponsored by a government, firm, or other non-academic entity?	80	69
1b. Eschew the use of placebo control and no treatment? (All Experiments, Field and Laboratory)	49	73
4. Eschew balancing (and/or stratifying) after experimentation?	80	88
11. Provide sufficient details to allow replication- including details on how and when treatments and controls were actually administered?	80	98

Source: All full-length articles that use randomization published in the *American Economic Review*, March 2000 – December 2009, excluding the *Proceedings*.

Note: "Percent Yes" is the total number of Yes responses divided by the relevant number of articles.

TABLE 3: DIFFERENCE IN RESULTS BETWEEN NEW ENGLAND JOURNAL OF MEDICINE AND AMERICAN ECONOMIC REVIEW SURVEY OF RANDOMIZATION, SIGNIFICANCE, AND VALIDITY (RANKED BY PERCENT YES, NEJM MINUS AER)

Survey Questions	Absolute Percent Difference
Does the study…	
4b. Does the study give estimated magnitudes of balance (or imbalance) in table, chart, or other form?	93
4a. Does the study discuss questions of balance for	84

sample strata and covariates?	
10. Defend the design of the experiment; for example, does it explain why it’s using artificial randomization? Or, does the study explain eligibility criteria for participants’ inclusion?	76
6. Test hypotheses (or treatments and controls) at multiple places?	62
9. Discuss its design relative to previous studies of the same?	59
20. Discuss the power of the test?	47
8. Does the study test the favored model on sub-populations or strata and/or test for robustness of estimates across different periods of time?	47
15. Explain the choice and implications of sample size?	43
13. Include cost and benefit data, such that one can determine the real net benefit of the experiment- adverse effects included?	42
25. Provide conflict of interest information?	39
22. Mimic conventional practice, in both experimentation and implementation, in the industry for the “treatment” in question?	31
3. Design balance covariates (confounding variables) prior to experimentation? Or, if the study is an observational study, does it stratify the sample prior to data collection?	26
24. Is the study sponsored by a government, firm, or other non-academic entity?	24
14. Include cost and benefit data such that the reader can estimate the net benefit of accepting the preferred hypothesis?	23
1b. Eschew the use of placebo control and no treatment? (All Experiments, Field and Laboratory)	22
2. Strive to test and estimate multiple treatments?	20
12. Compare results (such as effect sizes, magnitudes of coefficients, etc.) with previous studies?	12
5. Replicate a previous RCT or other types of empirical study?	10
7. Do diagnostic tests to assess “internal validity”?	5
7a. If the study does diagnostic tests to asses "internal validity" does it: do nothing (0), plots residuals for test of normally distributed error terms (1), or tests for independence of error terms (2)?	5

1a. Eschew the use of placebo control and no treatment? (All Field Experiments, not Laboratory)	4
11. Provide sufficient details to allow replication-including details on how and when treatments and controls were actually administered?	2
4. Eschew balancing (and/or stratifying) after experimentation?	1
17. Recognize that statistically insignificant results might have large and important substantive effects?	1
23. Consider the validity of other types of evidence, that is, evidence not based upon formal experiment and artificial randomization?	1
16. Acknowledge that statistical significance (or the lack of it) is not decisive?	0
18. Specify a loss function or otherwise stipulate the meaning of effect sizes, such that the reader can assess the how “large is large” and how “small is small” questions?	0
19. If it does not specify a loss function, does the study discuss otherwise the implications of being wrong about the preferred hypothesis?	0
21. Seek mainly to demonstrate substantive significance?	0

Note: “Absolute Percent Difference” is the absolute difference between the “Percent Yes” responses in the AER and the “Percent Yes” responses in the NEJM per question.

V. The Impartial Spectator against *Homo Experimentalis*

In a recent opinion piece published by Science, “*Homo Experimentalis* Evolves”, John List (2008) celebrates the randomized field experiments for which he is known.

> The fundamental challenge in the social sciences is how to go beyond correlational analysis to provide insights on causation. . . .Increasingly, insights on causation are also gained through the use of controlled experimentation. In this approach, causation is usually identified through randomization, much like controlled experiments used in drug trials. This approach combines the most attractive elements of the laboratory and of naturally occurring data: randomization.

The ethics of *homo experimentalis* are thus constrained by an erroneous conception of scientific method, and by an unexamined behaviorism. What would the impartial spectator say about randomized controlled trials? Testing the null hypothesis is not ethical if the randomized "control" is already known to be meaningfully less effective than best practice treatment. Justice requires best practice treatment, and proper benevolence should nudge it along. Prudence dictates due concern for efficiency and consistency, which have been shown to be lacking in randomized allocations of treatments and controls.

In Section III of The Theory of Moral Sentiments, "On Self Command," Smith observes (p. 280):

> The man who acts according to the rules of perfect prudence, of strict justice, and of proper benevolence, may be said to be perfectly virtuous. But the most perfect knowledge of these rules will not alone enable him to act in this manner: his own passions are very apt to mislead him; sometimes to drive him and sometimes to seduce him to violate all the rules which he himself, in all his sober and cool hours, approves of. The most perfect knowledge, if it is not supported by the most perfect self-command, will not always enable him to do his duty.

We have shown here that randomized controlled trials in economics (and in medicine, too) routinely violate the rules of perfect prudence, of strict justice, and of proper benevolence.

In medicine the alleged justification for randomized controlled trials is based on the idea of equipoise (Freedman,1987). Equipoise is a suspended state of belief, an "indifference" between two or more treatments. When equipoise exists, the mainstream medical community argues that a null hypothesis can be formed and tested using a randomized controlled trial (for example, Royall, 1991). "There is widespread agreement that ethics requires that each clinical trial begin with an honest null hypothesis" (Freedman, p. 141). We believe we have rejected the argument from equipoise and thus blind randomization, using theory and evidence illustrating Simpson's paradox (a big problem), Student's imbalance (a bigger problem), and ethical-economic thinking at the margin (the biggest problem of all).

At their best, randomized controlled trials are as David Cox noted preliminary studies, designed under desperation to alleviate pure ignorance and find something, anything. Deming agreed and noted that randomized studies offer no conclusions of real world importance, save possibly to help discover the categories and covariates to be stratified and balanced in a more relevant experiment. At their worst, randomized controlled trials are - like Tuskegee and Milgram – conclusive only in the moral stratum. On average, they do not filter out noise from signal.

In field experiments we are in a sense serving two different and powerful masters; the one, the warm glow in our hearts, our desire to help people in the real world, and to make a positive difference in their lives; the other, the cold and exclusively selfish tug for academic fame, wealth, and longevity. The one tug is active, the other passive, in its main appeal to our passions and reasons, affecting our self-command. Which master is recommended by the spectator is we believe now evident. Randomization + statistical significance = validity is a popular but false equation; the result is not equal to the cause and hasn't been since Student. Randomized trials are neither necessary nor sufficient. But they're often unethical.

Works Cited

Alderson, P. 1996. Equipoise as a means of managing uncertainty: Personal, communal, and proxy. Journal of Medical Ethics 22 (3, June): 135-139.

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

Angrist, J.D., and Pischke, J-S. 2009. Mostly harmless econometrics. Princeton: Princeton University Press.

Angrist, J.D., and Krueger, A.B. 2001. Instrumental variables and the search for identification: From supply and demand to natural experiments. Journal of Economic Perspectives 15 (4, Fall): 69-85.

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

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

Chalmers, I. 2003. Fisher and Bradford Hill: Theory and pragmatism? International Journal of Epidemiology 32: 922-924.

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

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

Cox, D. 1958. Planning of experiments. Wiley, New York.

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

DeMartino, G. 2011. The economist's oath: On the need for and content of professional economic ethics. New York: Oxford University Press.

Deming, W. E. 1978. Sample surveys: The field. In: Kruskal, W.H. and Tanur, J.M. (Eds.), International Encyclopedia of Statistics. The Free Press (Macmillan), New York and London, pp. 867-884.

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

Easterly, W. 2009. "The civil war in development economics," AidWatch: Just Asking That Aid Benefit the Poor. Blog of William Easterly: http://aidwatchers.com/2009/12/the-civil-war-in-development-economics/

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

Fairlie, R.W. and Robinson, J. 2013. Experimental evidence on the effects of home computers on academic achievement among schoolchildren. American Economic Review: Applied Economics 5 (3, July): 211-240.

Fehr, E. and Goette, L. 2007. Do workers work more if wages are high? Evidence from a randomized field experiment. American Economic Review 97 (March): 298-317.

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

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

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

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

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

Fisher, R.A., Yates, F. 1938. *Statistical Tables for Biological, Agricultural and Medical Research*. Edinburgh: Oliver and Boyd. Sixth edition.

Freedman, B. 1987. “Equipoise and the ethics of clinical research,” New England Journal of Medicine 317, No. 3: 141-145.

Glewwe, P., Park, A., and Zhao, M. 2012. Visualizing development: Eyeglasses and academic performance in primary schools in China. Center for International Food and Agricultural Policy Research, University of Minnesota, Working Paper WP12-2 (Jan.)

Gosset, W.S. 1936. Co-operation in large-scale experiments. Supplement to the Journal of the Royal Statistical Society 3: 115-36.

Gosset, W. S. 1962. Letters of William Sealy Gosset to R. A. Fisher. Vols. 1-5, Eckhart Library, University of Chicago. Private circulation.

Gray, F. D. 1998. The Tuskegee Syphilis Study. Montgomery: New South Books. http://www.cdc.gov/tuskegee/timeline.htm

Hall, A. D. 1905. The book of rothamsted experiments. E.P. Dutton and Company, New York.

Harrison, G. W. 2011. Randomization and its discontents. *Journal of African Economies* 20, 626-652.

Heckman, J.J., Vytlacil, E.J. 2007. Econometric evaluation of social programs, part I: causal models, structural models and econometric policy evaluation. In: Heckman, J.J., Leamer, E. (Eds.) Handbook of econometrics 6B. Elsevier, Amsterdam, pp. 4779-4874.

Herberich, D. H., S. D. Levitt, List, J.A. 2009. Can field experiments return agricultural economics to the glory days? American Journal of Agricultural Economics 91, 1259-1265.

Howick, J. 2011. The Philosophy of Evidence-Based Medicine. Oxford, UK: Wiley and Sons.

J-PAL Poverty Action Lab. 2010. When did randomized evaluations begin? Poverty Action Lab, MIT, Cambridge. http://www.povertyactionlab.org/methodology/when/when-did-randomized-evaluations-begin

Jayaraman, K. 1999. A Statistical Manual for Forestry Research. Bangkok: United Nations Food and Agriculture Organization.

Jensen, R.T., and Miller, N.H. 2008. Giffen Behavior and Subsistence Consumption. American Economic Review 98 (4):1553-1557.

Johnston, S. C., Rotenberg, J.D., Katrak, S., Smith, W.S., Elkins, J.S. 2006. Effect of a U.S. National Institutes of Health programme of clinical trials on public health and costs, Lancet 367: 1319-1327.

Jones, J. H. 1992. Bad Blood: The Tuskegee Syphilis Experiment. Free Press: 1992.

Kadane, J. B. 1986. Progress toward a more ethical method for clinical trials. The Journal of Medicine and Philosophy 11: 385-404.

Karlan, D., List, J. 2007. Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment. American Economic Review 97, 1774-1793.

Karlan, D., Appel, J. 2011. More than good intentions: How a new economics is helping to solve global poverty. Dutton, New York.

Lanham, R. 1991. A handlist of rhetorical terms. Los Angeles: University of California Press.

Levitt, S. D., List, J.A. 2009. Field experiments in economics: the past, the present, and the future. European Economic Review 53, 1-18.

Lilford, R. and Jackson, J. 1995. Equipoise and the ethics of randomization. Journal of the Royal Society of Medicine 88 (Oct.): 552-559.

Lindley, D. 1991. Making decisions. Wiley, New York.

List, J. 2008. Homo experimentalis evolves. Science 11 July 2008: Vol. 321 no. 5886 pp. 207-208.

McCloskey, D.N., Ziliak, S.T. 2010. Brief of *amici curiae* by statistics experts Professors Deirdre N. McCloskey and Stephen T. Ziliak in support of respondents, Vol. No. 09-1156, Supreme Court of the United States, Washington DC. Edward Labaton et al. Counsel of Record (Ed.), Matrixx et. al. v. Siracusano and NECA-IBEW Pension Fund, filed Nov. 12, 2010.

McCloskey, D.N. and Ziliak, S.T. 1996. The standard error of regressions. Journal of Economic Literature 34 (March): 97-114.

Mercer, W. B., Hall, A.D. 1911. The experimental error of yield trials. Journal of Agricultural Science 4, 107-127

Meyers, J., Sacks, G, van Es, H., Vanden Heuvel, J. 2011. Improving vineyard sampling efficiency via dynamic spatially explicit optimisation. Australian Journal of Grape and Wine Research 17, 306-315.

Milgram, S. 1974. Obedience to authority: An experimental view. New York: Harpercollins.

Mulligan, C. 2014. The economics of randomized experiments. The New York Times, Economix blog, March 5th, 2014. http://economix.blogs.nytimes.com/2014/03/05/the-economics-of-randomized-experiments/

Neyman, J., Iwaszkiewicz, K., Kolodziejczyk, S. 1935. Statistical problems in agricultural experimentation. Supplement to the Journal of the Royal Statistical Society 2: 107-80.

Neyman, J. 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.

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.

Ravallion, M. 2014. Two letters of correspondence to authors: comments on Ziliak's and Teather-Posadas's "The Unprincipled Randomization Principle in Economics and Medicine". Washington DC: Georgetown University, March 24th 2014.

Rawls, J. 1971. A theory of justice. Cambridge: Belknap Press.

Reddy, S. 2012. "Randomise this! On Poor Economics," Review of Agrarian Studies 2 (2): 60-73.

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.

Royall, R. M. 1991. Ethics and statistics in randomized clinical trials. Statistical Science 6 (1, Feb.): 52-62.

Rubin, D. 1990. Comment: Neyman (1923) and causal inference in experiments and observational studies. Statistical Science 5, 472–480.

Salazar-Lindo, E., Sebastian-Hart, J. et al. 2000. Racecadotril in the Treatment of Acute Watery Diarrhea in Children, New England Journal of Medicine 343 (7): 463-467.

Sen, A. 2006. What do we want from a theory of justice? The Journal of Philosophy 103 (5, May): 215-238.

Smith, A. 1759 [1791, 2009]. The theory of moral sentiments. New York: Penguin. Introduction by Amartya Sen.

Student. 1908. The probable error of a mean. Biometrika 6, 1-24.

Student. 1911. Appendix to Mercer and Hall's paper on 'The experimental error of field trials'. Journal of Agricultural Science 4, 128-131. In: Student. Student's collected papers, pp. 49-52. Pearson, E. and J. Wishart, (Eds.)

Student. 1923. On testing varieties of cereals. Biometrika 15, 271-293.

Student. 1926. Mathematics and agronomy. Journal of the American Society of Agronomy 18. Reprinted in: E. S. Pearson, Wishart, J. (Eds.), Student's Collected Papers (1942). University College London, London, pp. 121-34.

Student. 1938 [posthumous]. Comparison between balanced and random arrangements of field plots. Biometrika 29, 363-78.

Student. 1942 [posthumous]. Student's collected papers. University College London, London. E. S. Pearson, Wishart, J. (Eds.).

Thornton, R.L. 2008. The Demand for, and Impact of, Learning HIV Status. American Economic Review 98 (5): 1829-1863.

Tippett, L. 1958. The methods of experiments. Wiley, New York.

Varian, H. 2011. Are randomized trials the future of economics? Federalism offers opportunities for casual [sic] experimentation. The Economist, April 27th. http://www.economist.com/node/21256696

World Bank. 2004. Monitoring and evaluation: Some tools, methods, and approaches. Washington, DC: The World Bank.

Ziliak, S.T. 2014. Balanced versus randomized field experiments in economics: Why W.S. Gosset aka 'Student' matters. Review of Behavioral Economics 1 (1): 167-208.

Ziliak, S.T. 2011a. Matrixx v. Siracusano and Student v. 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. 2010a. The *Validus Medicus* and a new gold standard. The Lancet 376 (No. 9738, July): 324-325.

Ziliak, S.T. 2010b. Significant errors - Author's reply [to Stephen Senn], The Lancet 376 (No. 9750, Oct.): 1392.

Ziliak, S. T. 2008. Guinnessometrics: The economic foundation of 'Student's' *t*. Journal of Economic Perspectives 22 (Fall): 199-216.

Ziliak, S. T., McCloskey, D.N. 2008. The cult of statistical significance: How the standard error costs us jobs, justice, and lives. University of Michigan Press, Ann Arbor.