*Making Observations C O U N T**

*Presented to SOE, SUNY Albany

as part of Brown Bag series

22 September 2004

Bob Pruzek

Science entails ***careful observation***
and ***well-reasoned inference***

Scientific literacy is essential for all citizens
who aim to participate in society

Many excellent discussions of the role of science
are close at hand, on the web: *e.g.*
http://www.project2061.org/tools/sfaaol/chap1.htm
…and for more, just googlize!

My Main Goals are:

- To provide background and specific information, especially ***web sources & books that develop arguments supporting randomization*** in applied behavioral research
- To briefly *compare* ***evidence-based decisions*** *in* **health care** *with* ***evidence-based education****, noting some major similarities and important differences;* prospects for improvements in both areas will be briefly discussed, including an illustration
- To suggest some ***promising but mostly untried alternative research strategies & methods that use principles common to randomization, but aimed instead at observational studies***
- To ***motivate your study of methods*** that are *tightly organized around specific, carefully chosen questions*, especially as related to research about instruction, where ***as possible, 'likes are compared with likes'***

Randomized studies *(including randomized clinical trials)*

- ***What is randomization***? **ANS**: Assignment of individuals to two or more 'treatment' groups randomly, given an initial sample
- How can randomization be used? → in a wide variety of ways:
  - Broadly, as in random assignments of schools themselves, or classrooms to treatment and control sets;
  - Less broadly, as in assignment of students within classes to different treatments;
  - Narrowly, as in assignment within pairs or triplets (chosen to make them 'homogeneous') to respective treatments;

…and for the narrower of these think especially in terms of possibly ***short term treatments***…a few minutes, a half-hour, a class period …randomization methods can work well in a huge variety of situations, including those with short terms.

- Related to the last of the above-listed items, randomization is more effective if used ***within blocks*** **–** *e.g.*, homogeneous subgroups of schools, classes, students…as *two key advantages* follow from use of randomized blocks:

  * ***Efficiency***, as samples can often provide generalizable results even when they are not large

  * Enhanced possibilities for ***detecting interactions*** because treatments may work differently for different blocks; and learning specifics about interactions can be particularly helpful

- Most relevant to randomization studies in education is the notable book: ***Evidence Matters*** (2002): Mosteller & Boruch (eds), ISBN: 0-8157-0205-1 (pbk under $25.00). Chapter 1 is online at: www.brookings.org/dybdocroot/press/books/chapter_1/evidence_matters.pdf

You may wish to note reference to the ***Campbell Collaboration***, C2-SPECTR, a *registry* pertinent to several fields, including education, with FAQ-answers at: www.campbellcollaboration.org/Frafaqs.html

The ***Campbell Collaboration*** was inspired principally by the British counterpart, the ***Cochrane Collaboration***, founded in 1993, named after British epidemiologist, Archie Cochrane: see ***www.cochrane.org/index0.htm*** wherein its purpose is seen as ***providing*** "… ***accurate information about the effects of health-care readily available worldwide***." This collaboration produces and disseminates systematic reviews of ***health care interventions*** and promotes a search for evidence deriving from studies of interventions. ***Randomized clinical trials*** are given special emphasis, and more generally, the focus of collaborators is on ***evidence-based-medicine (EBM)*** to support recommendations for health care. The major product of this Collaboration is the ***Cochrane Database of Systematic Reviews*** (whence the idea for counterparts of these reviews in the ***Campbell Collaboration***)**.** Key accomplishments of EBM can be seen at sites such as: ***http://medicine.ucsf.edu/resources/guidelines/guide15.html#15***

However different the fields of medicine and health care, vs. education and educational practice, there are several similarities. In the past differences were emphasized, while recently more attention has been given to similarities. Notably, passionate and persistent arguments have been given by specialists to the effect that ***education should take its research model from medical research***, especially in the form of heavy reliance on randomized trials (*cf.* Mosteller and Boruch monograph, ***Evidence Matters***.)

The NO CHILD LEFT BEHIND ACT (Congress, 2002) institutionalizes this recommendation, with its strong & repeated emphasis on 's***cientifically based research' (SBR)** – which its authors see as* ***short-hand for research based on randomization***.

(For qualifications, but not outright dissent on this point, see Howard Gardner's (2002) essay in ***Education Week***: ***http://www.edweek.org/ew/newstory.cfm?slug=01gardner.h22*** and also the thoughtful essay by Hammersley (2001): ***http://www.leeds.ac.uk/educol/documents/00001819.htm***)

Whether the ***medical model*** can serve *centrally* as the model for educational research will not be debated here, but one cannot deny, the argument has gained a strong foothold. Various arguments for and against it are probably best seen in details, *i.e.,* particulars of the professional practice of education. This issue is likely to be worthy of your careful analysis.

Apart from arguing pros and cons of randomization, it may help more to focus on fundamentals: *viz.,*why is randomization so often advocated? In fact, this is easy to answer: It is because (in principle) ***randomization controls all forms of selection bias***. This means that, successfully executed, a *comparison of two or more randomly formed treatment groups entails* ***a scientifically impartial comparison for any outcome(s) one chooses to study****, the aim being to* ***eliminate selection bias***. *Authors of some methodological texts come close to arguing that without randomization, a study is basically flawed. Surely, this is too strong.*

Certainly it is ***naïve to believe that most questions in education are susceptible to randomized treatment comparisons***. While, physicians meet most of their patients only a few times each year, and restrict attention to a small number of topics, teachers meet their students most every day, often for many hours. Teachers are confronted with dozens of issues each day about which research has almost never been informing; moreover, environments of schools tend to deny all but a few prospects for using randomization – and often quite limited prospects for use of many other research methods. Although obvious, these points are rarely noted in methodological sources concerned with experimentation in education.

Dilemma*: if we aim for a scientific approach to research* ***we seek fairness in comparing treatments since we want to eliminate effects of selection bias; but we know that randomization is often simply an unrealistic means to achieve this end.***

Most interestingly, and providentially, we note that statisticians have recently developed ***new methods that can remove selection bias nearly as well as randomization – methods called*** ***<u>propensity score analysis</u> (PSA)***. PS methods help ***make groups comparable without the use of randomization***; that is, *PSA has been developed for (broad-scale) use in* ***observational research*** (*i.e., non*-experimental studies).

Consider some possible implications of this development: However appealing, randomized assignment to treatments are out of the question for many cogent research questions. The reasons are partly ***ethical***: one cannot reasonably ask some persons to smoke, others not; nor should we expect parents to consent to have their children assigned to what they deem to be the weaker instructional regimen, even if the researcher believes that superiority of the other has not been settled. But also, ***administrative (in)convenience***: Institutional Research Boards rarely approve proposals for experiments if there is any perceived risk; IRBs tend (understandably) to be conservative.

Although this is not the place for elaboration, it turns out that ***public health and medical research have seen many applications*** of PSA in the past dozen or so years, more so most recently. (Again, use *google* to elaborate on this – or ask me later.) Although PSA has seen few educational research applications, some forward momentum has begun to appear, as a web search will show.

Consider the following scenario to get a sense of how PSA might be used in an observational context.

Suppose you aim to research an issue concerning effects of different methods bearing on ***how children learn to read***. You have reviewed what good analysts believe to be central to the topic at hand; and your chosen theme is sufficiently general to be of broad interest, yet specific enough to center focused inquiry. You may have decided on phonemic awareness, phonics, (oral or silent) reading fluency, text comprehension or vocabulary; now ***you want to compare particular teaching methods that seem appealing***.

Suppose you approach your question(s) using ***an observational study, not an experiment***. What might you do? Suppose it is found that two or more teachers (probably at the same grade) with more or less comparable skills and experience, in two or more comparable schools, that ***have already chosen (or can be convinced to choose) two different ways to teach text comprehension or vocabulary*** (these two having been chosen quite arbitrarily). Consequently, the two methods are known to be 'appealing,' at least to these teachers. (We might help them 'refine' these teaching methods, and even monitor them over a term to ensure we can describe them.) We begin with the fact that there are two different, but (only) roughly comparable groups of students associated two instructional methods.

Use of ***PSA entails*** the following ***two steps*** (and logic):

1. Aim to observe – or collect records for – many ***covariates***: *variables that portray student (demographic/family) background, performance or achievements from the past, preferably most recent, etc., that are likely to* ***distinguish between children in these two groups***. (Several statistical methods come to mind as aiding this purpose [e.g., logistic regression], but just adding (standardized) achievement scores together might work well.) Based on derived variable 'X' we now *sort the children* (across the two groups) from high to low using variable X that distinguishes the teaching method groups. (Note that one need not make linearity assumptions, nor any other.)
2. Having sorted the children, we form subgroups (strata) that are ***homogeneous on X***, so that ***within subsets students should tend to be relatively similar with respect to ALL covariate distributions;*** then once the period of instruction has been completed, proceed to measure vocabulary and/or text comprehension skills of all children & ***compare student responses within strata***.

By having chosen to use covariates that distinguish between the two groups defined vis-a-vis teaching methods (that teachers themselves may have chosen) and then sorted children according to covariate information (X scores) into homogeneous subsets or strata, ***we will have removed, or at least adjusted for, most of the selection bias***. (PSA is the principal method that accomplishes this, and has also proven robust to methodological variations in its execution; conventional quasi-experimental methods do not do so.)

In the comparison of responses on outcomes of interest, the PS-based results may be as ***unassailable as those of an experiment, so that our results could be as interpretable as could be an experiment*** – supposing an experiment were possible. At times PSA's can even improve on randomization (since randomization can go awry). In effect, we have described how teachers may have 'performed an experiment' that an investigator has participated in through his/her organizational efforts and data collection. Such a methodology naturally has many possible variations, not unlike true experiments.

Consider what educational uses of PSA methodology could mean:

- We need ***not*** be concerned with getting ***consents*** from parents or school officials for student participation in the study – except for requesting sometimes that certain tests or surveys be administered.
- Definitions or specifications for teaching methods will almost necessarily be ***realistic***, as they are choices of what we usually can assume are experienced teachers; and we also may refine these.
- What we term 'treatments' (possibly instructional methods, but quite possibly qualitative or quantitative variations on methods) are now generally defined in particular ***contexts*** about which, as analysts, we shall become ***obligated to understand***.
- ***Design issues*** have mostly to do with what we choose to ***measure*** (covariates and responses), not how we allocate children to treatments – as the latter is irrelevant in observational studies.
- Finally, we still have latitude to ***define treatments narrowly or broadly***, much as with randomized studies. And we may still profit from use covariates to gain precision when comparing responses within strata.

Some final points:

- Fundamental to randomized comparisons as well as comparisons of individuals within propensity-defined strata is the notion of ***comparing 'likes with likes***.' The aim is to eliminate selection bias in either experiments or observational studies so as to aid interpretations. To the extent that one fails to compare 'likes with likes' there is a possibility that observed response similarities or differences can be explained on the basis of 'extraneous' (*prior*) differences on whatever covariates for which selection bias exists.
- As a goal, comparing 'likes with likes' is as old as argument itself; for we all know the phrase "Don't compare apples with oranges."
- To the extent to which we focus on comparison of 'likes with likes' we have reason to hope that for both covariates & responses, our ***observations will indeed count***.

NB: For a detailed account of propensity score analysis and graphics, go to www.albany.edu/~jz7088/559.html , click on '559', then go to Lecture 26 and click on the file you find there: RP & JH authors