Economics of Education Review 31 (2012) 744–748

Contents lists available at SciVerse ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Methods matter: Improving causal inference in educational and social science research: A review article

Eric R. Eide[a,*], Mark H. Showalter[b,1]

[a] Brigham Young University, Department of Economics, 130 FOB, Provo, UT 84602, United States
[b] Brigham Young University, Department of Economics, 179 FOB, Provo, UT 84602, United States

ARTICLE INFO

Article history:
Received 21 May 2012
Received in revised form 22 May 2012
Accepted 22 May 2012

JEL classification:
C1
I2

Keywords:
Causal estimation
Education research
Research design

ABSTRACT

Professors Richard J. Murnane and John B. Willett set out to capitalize on recent developments in education data and methodology by attempting to answer the following questions: How can new methods and data be applied most effectively in educational and social science research? What kinds of research designs are most appropriate? What kinds of data are needed? What statistical methods are best used to process these data, and how can results be interpreted so that policymakers are best informed? In this review we summarize main contributions of the book, assess the unique value-added of the text, and discuss the usability of it to various potential audiences.

1. Introduction

In recent decades, education researchers have witnessed a marked improvement in the quantity and quality of data, in research design, and in statistical methodology that can lead to credible causal inferences. Murnane and Willett (2011) set out to capitalize on these recent developments by attempting to answer the following questions:

- How can new methods and data be applied most effectively in educational and social science research?
- What kinds of research designs are most appropriate?
- What kinds of data are needed?
- What statistical methods are best used to process these data, and how can results be interpreted so that policymakers are best informed?

The result of these efforts is an incredibly useful book that will serve as a valuable resource to researchers and students interested in education and social science research. In this review we summarize main contributions of the book and highlight content we think is particularly relevant for economists. We also assess the unique value-added of the text and discuss the usability of it to various potential audiences.

2. Content summary

In this section we summarize the authors' discussion of each topic with an emphasis on the empirical examples. Our intent is to provide a glimpse into how the authors' approach the material covered in the book.

2.1. Background

Murnane and Willett begin by motivating the importance of causal estimation when answering education questions. They emphasize theory's role in guiding empirical research and in measuring key conceptual variables

* Corresponding author. Tel.: +1 801 422 2859; fax: +1 801 422 0194.
E-mail addresses: eide@byu.edu (E.R. Eide), showalter@byu.edu (M.H. Showalter).
[1] Tel.: +1 801 422 4645; fax: +1 801 422 0194.

0272-7757/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.econedurev.2012.05.010

relevant to the researcher's hypothesis. In the economics of education the underlying theory is often implied rather than explicitly stated, for example the human capital model, but it is nevertheless important to ground the empirical work in theory.

A strength of the book is that it illustrates concepts with relevant high quality studies. As an example of a randomized experimental design they rely on the Tennessee Student/Teacher Achievement Ratio (STAR) experiment (Krueger, 1999), which involved more than 11,000 students and 1300 teachers in 79 public schools throughout the state. Children starting kindergarten in the fall of 1985 were randomly assigned to one of the following three types of classes: a small class with 13–17 students, a regular size class with 22–25 students, or a regular size class with a full-time teacher's aide. Teachers in each school were also randomly assigned to classrooms, and students were to remain in their originally assigned class type through third grade. Because of its randomized experimental design, the Tennessee STAR project is used throughout the book to illustrate different issues that arise in obtaining causal estimates from an educational intervention.

2.2. Design

In an experimental design, researchers would ideally like to observe and compare the outcomes of each child under two different yet concurrent conditions: the treatment condition and the control or counterfactual condition. While we cannot observe the *individual* treatment effect because once a child is assigned to either treatment or control we cannot observe their outcome under the other state, we can estimate the *average* treatment effect (ATE). The ATE across all children in the population is the population average or expectation of the individual treatment effects. An important nuance of this discussion is that members of the treatment and control groups are *equal in expectation*. Potential members of the treatment and control groups would be identical on all observed and unobserved characteristics, on average, in the population.

To illustrate important principles of designing randomized experiments, the authors use data from the School Choice Scholarships Foundation (SCSF) initiative which provides scholarships of up to $1400–1300 children from low-income families who were enrolled in New York City public schools. These scholarships, which are renewable for up to three years, can be used to pay for tuition at religious or secular private schools. A lottery was used to award scholarships to the more than 10,000 applicants. The authors refer to this initiative as the New York Scholarship Program (NYSP), and it is used frequently throughout the text.

While random assignment experiments are often considered the "gold standard," the authors are careful to note the potential problems that can arise in their design and implementation. The discussion focuses on the following decisions: defining the treatment; defining the population from which participants will be sampled; deciding which outcomes to measure; and deciding how long to track participants. Even with careful thought about the design of the experiment, there are a number of potential threats to internal and external validity. Threats include contamination of treatment-control contrast; cross-overs of participants between the control and treatment group after random assignment has taken place; sample attrition; and changes in participants' behavior due to participation in the experiment itself. This discussion reminds researchers about the various ways in which causal estimates can be compromised. Careful researchers will take these points into consideration.

To demonstrate the decisions encountered in designing and implementing randomized experiments, the authors discuss a study by MDRC, a contract research firm, which examined the educational impact of career academies. As described by MDRC, career academies aim to restructure large high schools into smaller learning communities and create improved transitions from high school into the workplace and further education (http://www.mdrc.org/project_29_1.html). They enroll 30–60 students per grade and are organized around themes such as health sciences, law, business and finance, and pre-engineering. Students who applied for admission to the career academies were selected by a lottery. Researchers measured students' educational outcomes (test scores, high school graduation rates, and college enrollment rates), labor market outcomes (employment rates, and earnings), and family outcomes (marriage and child support rates) and compared them between treatment and control groups.

2.3. Statistical power and sample size

The authors review the role of sample size in detecting causal effects. They work through how to conduct a power analysis, including the software tools that are available. As part of the discussion they highlight the strengths and limitations of parametric tests, including the role of assumptions when adding more structure to the estimation approach, the usefulness of covariates in adding statistical power, the reliability of outcome measures (measurement error), and the choice between one-tailed and two-tailed tests.

2.4. Experimental research when participants are clustered within intact groups

In education research sometimes intact groups are randomized into treatment and control groups rather than the individual units. For example, intact schools rather than individual students may be the unit that is randomized. The authors discuss two ways to handle randomized experiments when participants are clustered within intact groups: random effects and fixed effects.

Random and fixed effects are discussed in the context of a randomization of intact schools to the Success for All (SFA) program, a school-wide intervention with the goal to help every student in a school to perform at grade level in reading by the end of third grade. Murnane and Willett's analysis of the effectiveness of this randomized experiment of intact schools found that students who were part of the SFA treatment group have substantially higher scores on a reading measure, with an effect size of just over one-fifth of