



education review // reseñas educativas

editors: gene v glass gustavo e. fischman melissa cast-brede

a multi-lingual journal of book reviews

November 23, 2011

ISSN 1094-5296

Education Review/Reseñas Educativas is a project of the National Education Policy Center <http://nepc.colorado.edu>

Follow Education Review on Facebook

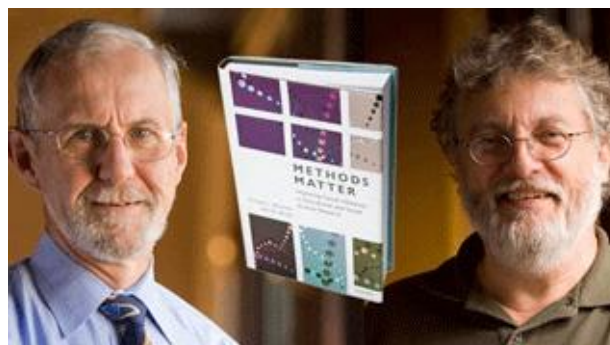


Murnane, Richard J. & John B. Willett. (2010) *Methods Matter: Improving Causal Inference in Educational and Social Science Research*. NY: Oxford University Press.

Pp. 416 ISBN 978-0199753864

Reviewed by Gregory Camilli
University of Colorado Boulder

Methods Matter, by Richard Murnane and John Willett, can be conceptually divided into three parts: discussion of statistical methods for obtaining unbiased estimates of causal effects; use of education studies to illustrate causal methods; and a set of recommendations ensuing from the research literature employing causal methods. For the purpose of this review, it is first necessary to parse the phrase “causal inference.”



Richard J. Murnane

John B. Willett

Citation: Camilli, Gregory. (2011 November 23) Review of *Methods Matter* by Richard J. Murnane & John B. Willett. *Education Review*, 14. Retrieved [Date] from <http://www.edrev.info/reviews/rev1127.pdf>

As I understand it, researchers using causal analysis methods do two different things: they obtain unbiased estimates of causal effects, and subsequently, they make inferences more or less supported by these estimates. Unsurprisingly, the former is much easier because the latter involves a host of unknown factors. This might strike the student of causal methods as incorrect... aren't these methods devised precisely to eliminate such unknown effects? No, actually, they are not. As Murnane and Willet explain

Randomized experiments and quasi-experiments typically provide estimates of the total effect of a policy intervention on one or more outcomes, not the effects of the intervention holding constant the levels of other inputs.... (p. 31)

A fancier way to say this is that there may be substantial endogeneity of causal mechanisms within a treatment group, and this endogeneity gets "absorbed" by the treatment. In simpler terms, one could say that the *treatment* is an umbrella for a number of influences on outcomes, and some of these influences may be highly interactive. There are numerous examples in *Methods Matter* that should serve as cautions to those who believe randomized experiments are the "gold standard" for making causal inferences. Berk (2005) characterized true experiments more modestly as the "bronze standard." I will return to this point in more detail when considering the recommendations offered by the authors.

Having made the distinction, each of the three conceptual divisions are described below. The first two divisions are not contiguous in the pages of the book, but various recommendations and conclusions are offered in the last two chapters.

The authors inform the reader that the "new methods of causal inference" had evolved over the last 15 years, so it seems wise to facilitate the implementation of these methods—especially because existing methods could not "support credible statements of cause and effect" (p. xiii). I was eager to read something new in terms of methods, but it soon became clear that the "new methods" stretched back over a period of about 35 years (dated roughly to the development of Donald Rubin's potential outcomes model;

see Rosenbaum & Rubin, 1984). One could point out that Rubin's model was presaged by the selection model of Roy (1951) as noted acerbically by Heckman (2006), or even the development of the instrumental variables methods in 1928 (Stock & Trebbi, 2003). Statisticians—especially econometricians—do disagree about the ancestry of these methods, but the point is that conceptual development of causal methods has been underway for some time. Indeed, previous books such as *Counterfactuals and Causal inference* (Morgan & Winship, 2007) cover much of the material presented here in a more detailed manner. What is unique and *Methods Matter* is the focus on education research by economists on an international scale. It is virtually canon in the book to present rich empirical examples for illustrating points that are made more abstractly in other sources. Some material on power analysis in cluster randomized studies and regression discontinuity designs is newer, as are the theories of LATE estimators (Angrist, Imbens, & Rubin, 1996). In general, I found the material on instrumental variables (Chapters 10 and 11) to be exceptionally well-written. The materials on propensity score matching (Rosenbaum & Rubin, 1983), and combining different methods, such as regression discontinuity and second-order differencing, is also newer. Other methods such as OLS regression (and bias in OLS coefficients), mean comparisons via *t*-test, stratification, and covariance analysis are age-old. From the corpus of all of this material, including cogently explained rationales for different methods, it becomes clear that the design of studies to estimate causal effects is a craft, not a science and even less a set of technical skills. Indeed, qualitative reasoning (summarized in Chapter 13) should be considered mandatory for intelligent research design.

Along with rich empirical examples of estimation methods, Murnane and Willett also attempt to walk the reader didactically through elementary statistical analysis as well as issues in the design of experiments. For example, on pages 52-54 a brief description is given for the frequentist model of the standard error, followed by the elements of power analysis on pages 83-92 (prior to the examination of intraclass correlations and power for cluster experiments). Certainly this material is prerequisite to a full appreciation of the material in the book, but one or two semesters of study in statistical methods may be

necessary for a full appreciation. For readers with sufficient background, this material may serve as an effective review. In any case, *Methods Matter* would serve well as one resource for a class on experimental and quasi-experimental design. For example, one could make use of the material on power analysis to have students estimate the power to detect an effect size of .2 in the first-year SFA data analysis (see pp. 122-128).

I am less sanguine about the substantive lessons of the book, offered in Chapter 14. The authors present four guidelines (pp. 351-352) for policy emanating from evidence amassed from newer methods of estimating causal effects:

1. "School enrollment decisions are very sensitive to cost, especially for low-income families."
2. "a necessary condition for improving school quality and student achievement is that the daily experience of students must change."
3. "providing explicit incentives for teachers and students can be a powerful way of improving student outcomes for students."
4. "policies to increase the schooling options available to children from poor families have significant promise."

While the estimates of causal effects presented in the book appear to be consistent with these recommendations, I had no sense that the authors attempted to illustrate the broader question of causal inference. For example, under guideline 4 the authors refer to evidential support from a study carried out in rural India by the Abdul Latif Jameel Poverty Action Lab (J-PAL), where teacher absenteeism is a significant problem. Here, nongovernmental organizations (NGO) train teachers, who typically have a high school education and are paid on short-term contracts at "quite low rates of pay." An experiment was designed for the NGO (Seva Mandir) to estimate the effect of basing teacher pay on the number of days actually taught, rather than the typical flat monthly contract. The control group was paid approximately 1000 rupees per month, and the treatment group 50 rupees for classroom-present day to a maximum of 1,300 rupees per month. From the 27-month experiment, a 21% decrease in absenteeism was found (from 42% in the control group to 21% in the treatment

group). Teachers were counted as present based on “tamper-proof” classroom cameras (to be present, two photos of a teachers were required—five hours apart—on a given day with a minimum number of students). In Dufo and Hanna (2006), more detail is given. Teachers’ base salary was Rs1000 per month (about \$23). In the treatment schools, teachers received Rs50 (about \$1.15) if they attended in excess of the 20 days, and they received a Rs50 fine for each of the 20 days absent. The maximum salary was capped at RS1300, and the maximum fine was capped at Rs500.

Though the program was reported to be a success, a number of subsequent investigations of causal mechanism might have strengthened this claim. Assuming that schools were open 6 days a week, a teaching month would be about 27 days. So an incentivized salary of about Rs1300 was possible in the treatment group. What, in fact, was the average salary in the treatment group? Answer: about 1000Rs. On average, the number of teaching days increased; but the incentives and fines apparently cancelled. So how did this work? Depending on one’s point of view, this question about the causal mechanism is either idle curiosity, or is necessary for creating a sustainable program. Second, the cost of the program (mainly camera equipment) was Rs5379 (about \$120) per teacher per year, or about 40% of the average teacher salary. It is doubtful that such a program would be sustainable without a steep drop in the price of camera technology. Third, it would seem that more information was available for examining the causal mechanism, because teachers with higher rates of classroom attendance and higher salaries presumably should have had higher student outcomes.

Another example concerns vouchers. Murnane and Willett cite Angrist, Bettinger, and Kremer (2006) as demonstrating the success of a voucher program in Columbia. They also cite Rouse (1998) that the voucher program in Milwaukee led to improvement in mathematics, but not reading. With all due respect, the latter study is fairly old and newer information is readily available. For example, Witte, Carlson, Cowen, Fleming, and Wolf (2011) report that:

The primary finding in all of these comparisons is that there are no statistically significant

differences in student achievement growth in either math or reading between MPCP and MPS students three years after they were carefully matched to each other. (p. ii)

Similar findings were available in reports by the same authors in 2009 and 2010. I would summarize that in some context voucher programs have been effective, but in others, they clearly have not. It makes no sense to estimate causal effects if there is no plan to investigate the causal mechanisms more broadly—especially those that lead to differential outcomes across contexts. Cronbach was clearly the master of tackling the complexity of the endogeneity in social programs. His descriptive methods in *Designing Educational and Social Programs* (especially the UTOS framework) could also be described as craft, if not lost art, and might provide a more fundamental understanding of causal inference if paired with *Methods Matter*.

Murnane and Willett believe there are “challenges to answering descriptive questions well,” they also assert that “these challenges are typically less difficult than the challenges you will face with causal questions.” I would argue that estimating effects is something that quantitative researchers do well, but proficiency in causal inference is often downplayed, perhaps because contextual reasoning is messy, laborious, and is aided by training in qualitative methods, e.g., case study or interviewing. Estimating causal effects is valuable, but estimation is the easy part. Causal effects, whatever their value, are effects. Presumably, the wisdom inherent in the experimental manipulation goes part way in determining the causes of those effects; but as seen in the J-PAL study, human responses to being “treated” are often vastly more complex than a simple *a priori* description of treatment.

To be fair, a text on causal methods would not be expected to undertake the kind of detail provided by Cronbach (1982) on the INCAP study (pp. 46-59), yet it is still proper to represent the role of *descriptive research*, both quantitative and qualitative, in an accurate light that will facilitate professional development. Murnane and Willett wrote:

the phrase ‘high-quality’ itself eventually becomes code for referring to studies that

effectively employ the approaches we describe.
(p. 12)

I agree that high quality studies are useful, but I would add that it would be better to assert that a carefully causal design is code for high-quality because other orientations should also be recognized as having this potential. Moreover, the application of causal methods does not necessarily lead to high-quality inferences, and it is difficult to fathom the existence of indisputable technical information. Rather, applying the information obtained from high quality studies presents:

Challenges that require thoughtful analyses with limited or even insufficient information. That require effective responses within ambiguous circumstances or complex economic and political contexts. That, most of all, demand decisive action that must be articulated—and even defended—among other talented, ambitious individuals. (p. 3)

Some few readers might recognize this short quotation as the rationale for the Harvard Business School's pedagogical method of case-based learning: <http://www.hbs.edu/learning/case.html>. With studies designed to facilitate causal inferences, I would argue that we can do one step better in reducing the ambiguity inherent in generalizing causal effects by studying the heterogeneity of effects across different populations and context. A systematic review (Higgins & Green, 2011) of related causal studies would strengthen claims made under the four guidelines above.

Though I disagree with the orientation of Murnane and Willett on some matters, as noted above, I do believe that *Methods Matter* is a useful book due to its richly empirical documentation. I would give this book strong consideration as a text in a class on the design of quantitative studies. It serves as a useful bridge between the estimation of causal effects and the development of causal inferences. However, the bridge needs to be crossed, and instructors will recognize that success in this matter requires a substantial amount of effort as well as complementary materials.

References

- Angrist, J., Bettinger, E., & Kremer, M. (2006). Long-term educational consequences of secondary school vouchers: Evidence from administrative records in Colombia. *American Economic Review*, 96(3), 847-862.
- Angrist, J.D., Imbens, G.W., & Rubin, D.B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91, 444-455.
- Berk, R. (2005). Randomized Experiments as the Bronze Standard. UC Los Angeles: Department of Statistics, UCLA. Retrieved 11/16/2011 from: <http://escholarship.org/uc/item/2d57d1kq>.
- Cronbach, L.J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.
- Duflo, E. & Hanna, R. (2006). Monitoring works: Getting teachers to come to school. Natural Field Experiments 0038, The Field Experiments Website at <http://ideas.repec.org/s/feb/natura.html>.
- Heckman, J.J. (2006), Rejoinder: Response to Sobel. *Sociological Methodology*, 35(1), 135-150.
- Higgins, J.P.T. & Green, S. (editors). *Cochrane handbook for systematic reviews of interventions*, Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.
- Morgan, S., and Winship, C. (2007). *Counterfactuals and causal Inference: Methods and principles for social research*. Cambridge University Press.
- Rosenbaum, P.R. & Rubin, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516-524.
- Rouse, C.E. (1998). Private school vouchers and student achievement: an evaluation of the Milwaukee Parental

Choice Program. *The Quarterly Journal of Economics*, 113(2), 553-602.

Roy, A. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3, 135-46.

Stock, J. H. & Trebbi, F. (2003). Retrospectives: Who invented instrumental variable regression? *Journal of Economic Perspectives*, 17(3), 177-194.

Witte, John; Patrick Wolf; Joshua Cowen & David Fleming. (2011, March). Milwaukee Parental Choice Program: Longitudinal Education Growth Study Fourth Year Report. School Choice Demonstration Project: Milwaukee Evaluation, Year 4.
http://www.uark.edu/ua/der/SCDP/Milwaukee_Eval/Report_23.pdf

About the Reviewer

Gregory Camilli
Professor, School of Education
University of Colorado Boulder

Gregory Camilli received his PhD from the Research and Evaluation Methodology Program at the University of Colorado Boulder. Prior to returning to CU, he spent 23 years at Rutgers, State University of New Jersey. In addition to teaching classes in statistics, measurement, and meta-analysis, his research interests have focused on early education interventions, school factors in mathematics achievement, and multilevel item response theory (IRT). He recently completed a term as Co-Editor of *Educational Researcher* and serves as an Associate Editor of the *Journal of Educational and Behavioral Statistics*. His research has appeared in the *Teachers College Record*, *Reading Research Quarterly*, *Journal of Educational Measurement*, *Journal of Educational and Behavioral Statistics*, and *Educational Measurement, Fourth Edition*. His current research projects include the effects of affirmative action in law school admissions, the use of multilevel IRT modeling for identifying items are to instructional effects, and Bayesian regression for meta-analysis. His research has been funded by the National Science Foundation, the Law School Admission Council, the Pew Charitable Trusts, and the National Center for Early Education Research.



Copyright is retained by the first or sole author, who grants right of first publication to the *Education Review*.

Education Review/Reseñas Educativas is a project of the National Education Policy Center <http://nepc.colorado.edu>



Editors

Gene V Glass

glass@edrev.info

Gustavo Fischman

fischman@edrev.info

Melissa Cast-Brede

cast-brede@edrev.info
