# Rubin Causal Model and Selection

## Broad References

Books

– Imbens, G. W., Rubin, D. B. (2015)***. *Causal Inference in Statistics, Social, and Biomedical Sciences.* Cambridge University Press. Chapters 1-4.

– Angrist, J. D., Pischke, J. S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton university press.

Papers

– Holland, P. W. (1986). "Statistics and Causal Inference"*. *Journal of the American Statistical Association*, 81(396), 945-960.

> Problems involving causal inference have dogged at the heels of statistics since its earliest days. Correlation does not imply causation, and yet causal conclusions drawn from a carefully designed experiment are often valid. What can a statistical model say about causation? This question is addressed by using a particular model for causal inference (Holland and Rubin 1983; Rubin 1974) to critique the discussions of other writers on causation and causal inference. These include selected philosophers, medical researchers, statisticians, econometricians, and proponents of causal modeling.

– Todd, P. E., Wolpin, K. I. (2003)*** "On the Specification and Estimation of the Production Function for Cognitive Achievement". *The Economic Journal*, 113(485), F3-F33.[1], specially section 2.3.4.

> This paper considers methods for modelling the production function for cognitive achievement in a way that captures theoretical notions that child development is a cumulative process depending on the history of family and school inputs and on innate ability. It develops a general modelling framework that accommodates many of the estimating equations used in the literatures. It considers different ways of addressing data limitations, and it makes precise the identifying assumptions needed to justify alternative approaches. Commonly used specifications are shown to place restrictive assumptions on the production technology. Ways of testing modelling assumptions and of relaxing them are discussed.

*I believe this is a canonical paper on the topic. The most important part is the second half, from section 2.2. onwards, equation (3) is the model proposed in its most general form, where cognitive achievement, as measured by test performance at some particular age, is the outcome of a cumulative process of knowledge acquisition that combines the whole history of family inputs, school inputs, genetic endowments and random error. The two basic problems in estimating such a model are omitted variables (genetic endowments) and missing data on the whole history of inputs. Section 2.3. presents an "inventory" of the specifications in the literature, and the way they deal with these problems (see Table 3). Importantly, the author is explicit about assumptions on the production technology and on the input decision rules that would justify its applications. The inventory encompasses the "contemporaneous specificication", "value added", "cumulative". Some takeaways from the basic "value added" model are very interesting, for instance that any model that admits the presence of unobserved endowments must also deal with the fact that this will be most likely correlated with baseline achievement, and this would bias* all *estimates in the regression. Importantly, section 2.3.4. is a discussion on teacher and school value added estimates: are they policy effects or production function parameters? This is an important discussion for the interpretation of these estimates. This is an overarching topic of the paper.*

– Duflo, E., Glennerster, R., Kremer, M. (2007)**. "Using Randomization in Development Economics Research: A Toolkit". *Handbook of Development Economics*, 4, 3895-3962, Sections 2 and 8.

[1] In general, this is a paper that seems worth studying much more carefully than I was able to do.

This paper is a practical guide (a toolkit) for researchers, students and practitioners wishing to introduce randomization as part of a research design in the field. It first covers the rationale for the use of randomization, as a solution to selection bias and a partial solution to publication biases. Second, it discusses various ways in which randomization can be practically introduced in a field settings. Third, it discusses designs issues such as sample size requirements, stratification, level of randomization and data collection methods. Fourth, it discusses how to analyze data from randomized evaluations when there are departures from the basic framework. It reviews in particular how to handle imperfect compliance and externalities. Finally, it discusses some of the issues involved in drawing general conclusions from randomized evaluations, including the necessary use of theory as a guide when designing evaluations and interpreting results.