Limits of Econometrics

David A. Freedman®

1. INTRODUCTION

It is an article of faith in much applied work that disturbance terms are IID—Independent and Identically Distributed—across observations. Sometimes, this assumption is replaced by other assumptions that are more complicated but equally artificial. For example, when observations are ordered in time, the disturbance terms ε_t are sometimes assumed to follow an "autoregression," e.g., $\varepsilon_t = \lambda\, \varepsilon_{t-1} + \delta_t$, where now λ is a parameter to be estimated, and it is the δ_t that are IID. However, there is an alternative that should always be kept in mind. Disturbances are DDD—Dependent and Differently Distributed—across subjects. In the autoregression, for example, the δ_t could easily be DDD, and introducing yet another model would only postpone the moment of truth.

A second article of faith for many applied workers is that functions are linear with coefficients that are constant across subjects. The alternative is that functions are non-linear, with coefficients (or parameters more generally) that vary across subjects. The dueling acronyms would be LCC (Linear with Constant Coefficients) and NLNC (Non-Linear with Non-constant Coefficients). Some models have "random coefficients", which only delays the inevitable: coefficients are assumed to be drawn at random from distributions that are constant across subjects. Why would that be so?

These articles of faith have had considerable influence on the applied literature. Therefore, when reading a statistical study, try to find out what kind of statistical analysis got the authors from the data to the conclusions. What are the assumptions behind the analysis? Are these assumptions plausible? What is allowed to vary and what is taken to be constant? If causal inferences are made from observational data, why are parameters invariant under interventions? Where are the response schedules? Do the response schedules describe reasonable thought experiments?

For applied workers who are going to publish research based on statistical models, the recommendation is to archive the data, the equations, and the programs. This would allow replication, at least in the narrowest sense of the term (Dewald et al., 1986; Hubbard et al., 1998). Assumptions should be made explicit. It should be made clear which assumptions were checked, and how the checking was done. It should also be made clear which assumptions were not checked. Stating the model clearly is a good first step—and a step which is omitted with remarkable frequency, even in the best journals.

Modelers may feel there are responses to some of these objections. For example, a variety of techniques can be used when developing a model, including regression diagnostics, specification tests, and formalized model selection procedures. These techniques might well be helpful. For instance, diagnostics are seldom reported in applied papers, and should probably be used more often.

® Revised version of Chapter 8 in David A. Freedman (2005). Statistical Models: Theory and Practice. Cambridge University Press.

In the end, however, such things work only if there is some relatively localized breakdown in the modeling assumptions—a technical problem which has a technical fix. There is no way to infer the "right" model from the data unless there is strong prior theory to limit the universe of possible models. More technically, diagnostics and specification tests usually have good power only against restricted classes of alternatives (Freedman, 2008). The kind of strong theory needed to restrict the universe of models is rarely available in the social sciences.

Model selection procedures like AIC (Akaike's Information Criterion) only work—under suitable regularity conditions—"in the limit," as sample size goes to infinity. Even then, AIC overfits. Therefore, behavior in finite samples needs to be assessed. Such assessments are unusual. Moreover, AIC and the like are commonly used in cases where the regularity conditions do not hold, so operating characteristics of the procedures are unknown, even with very large samples. Specification tests are open to similar objections.

Bayesian methods are sometimes thought to solve the model selection problem (and other problems too). However, in non-parametric settings, even a strictly Bayesian approach can lead to inconsistency, often because of overfitting. "Priors" that have infinite mass or depend on the data merely cloud the issue. For reviews, see Diaconis and Freedman (1998), Eaton and Freedman (2004), Freedman (1995).

2.1. The Bootstrap

How does the bootstrap fit into this picture? The bootstrap is in many cases a helpful way to compute standard errors—*given* the model. The bootstrap usually cannot answer basic questions about validity of the model, but it can sometimes be used to assess impacts of relatively minor failures in assumptions. The bootstrap has been used to create chance models from data sets, and some observers will find this pleasing.

2.1. The Role of Asymptotics

Statistical procedures are often defended on the basis of their "asymptotic" properties—the way they behave when the sample is large. See, for instance, Beck (2001:273): "methods can be theoretically justified based on their large-[sample] behavior." This is an over simplification. If we have a sample of size 100, what would happen with a sample of size 100,000 is not a decisive consideration. Asymptotics are useful because they give clues to behavior for samples like the one you actually have. Furthermore, asymptotics set a threshold. Procedures that do badly with large samples are unlikely to do well with small samples.

With the central limit theorem, the asymptotics take hold rather quickly: when the sample size is 25, the normal curve is a often a good approximation to the probability histogram for the sample average; when the sample size is 100, the approximation is often excellent. With feasible GLS, on the other hand, if there are a lot of covariances to estimate, the asymptotics take hold rather slowly (Freedman 2005, chapter 7).

The difficulties in modeling are not unknown. For example, Hendry (1980:390) writes that "Econometricians have found their Philosophers' Stone; it is called regression analysis and is used for transforming data into 'significant' results!" This seriously under-estimates the number of philosophers' stones. Hendry's position is more complicated than the quote might suggest. Other responses from the modeling perspective are quite predictable.

Philosophers' stones in the early twenty-first century

> *Correlation, partial correlation, Cross lagged correlation, Principal components, Factor analysis, OLS,GLS, PLS, IISLS, IIISLS, IVLS, FIML, LIML, SEM, GLM, HLM, HMM, GMM, ANOVA, MANOVA, Meta-analysis, Logits, Probits, Ridits, Tobits, RESET, DFITS, AIC, BIC, MAXENT, MDL, VAR, AR, ARIMA, ARFIMA, ARCH, GARCH, LISREL, Partial likelihood, Proportional hazards, Hinges, Froots, Flogs with median polish, CART, Boosting, Bagging, MARS, LARS, LASSO, Neural nets, Expert systems, Bayesian expert systems, Ignorance priors, WinBUGS, EM, LM, MCMC, DAGs, TETRAD, TETRAD II....*

The modelers' response

> *We know all that. Nothing is perfect. Linearity has to be a good first approximation. Log linearity has to be a good first approximation. The assumptions are reasonable. The assumptions don't matter. The assumptions are conservative. You can't prove the assumptions are wrong. The biases will cancel. We can model the biases. We're only doing what everybody else does. Now we use more sophisticated techniques. If we don't do it, someone else will. What would you do? The decision-maker has to be better off with us than without us. We all have mental models. Not using a model is still a model. The models aren't totally useless. You have to do the best you can with the data. You have to make assumptions in order to make progress. You have to give the models the benefit of the doubt. Where's the harm?*

2. CRITICAL LITERATURE

For the better part of a century, many scholars in many different disciplines have expressed considerable skepticism about the possibility of disentangling complex causal processes by means of statistical modeling. Some of this critical literature will be reviewed here. The starting point is the exchange between Keynes (1939, 1940) and Tinbergen (1940). Tinbergen was one of the pioneers of econometric modeling. Keynes expressed blank disbelief about the development:

> "No one could be more frank, more painstaking, more free from subjective bias or *parti pris* than Professor Tinbergen. There is no one, therefore, so far as human qualities go, whom it would be safer to trust with black magic. That there is anyone I would trust with it at the present stage, or that this brand of statistical alchemy is ripe to become a branch of science, I am not yet persuaded. But Newton, Boyle and Locke all played with alchemy. So let him continue" (Keynes 1940:156).

Other familiar citations in the economics literature include Liu (1960), Lucas (1976) and Sims (1980). Lucas was concerned about parameters that changed under intervention. Manski (1995) returns to the problem of under-identification that was posed so sharply by Liu (1960) and Sims (1980): in brief, a priori exclusion of variables from causal equations can seldom be justified, so there will typically be more parameters than data. Manski suggests methods for bounding quantities that cannot be estimated. Sims' idea was to use low-dimensional models for policy analysis, instead of complex high-dimensional ones. Leamer (1978) discusses the issues created by inferring specifications from the data, as does Hendry (1980). Engle et al. (1983) distinguish several kinds of exogeneity assumptions.

Heckman (2000) traces the development of econometric thought from Haavelmo and Frisch onwards. Potential outcomes and structural parameters play a central role, but "the empirical track record of the structural [modeling] approach is, at best, mixed" (p. 49). Instead, the fundamental contributions of econometrics are the insights

> "that causality is a property of a model, that many models may explain the same data and that assumptions must be made to identify causal or structural models..." (p. 89).

Moreover, econometricians have clarified "the possibility of interrelationships among causes," as well as "the conditional nature of causal knowledge and the impossibility of a purely empirical approach to analyzing causal questions" (pp. 89–90). Heckman concludes that

> "The information in any body of data is usually too weak to eliminate competing causal explanations of the same phenomenon. There is no mechanical algorithm for producing a set of 'assumption free' facts or causal estimates based on those facts" (p. 91).

Some econometricians have turned to natural experiments for the evaluation of causal theories. These investigators stress the value of strong research designs, with careful data collection and thorough, context specific, data analysis. Angrist and Krueger (2001) have a useful survey.

Rational choice theory is a frequently-offered justification for statistical modeling in economics and cognate fields. Therefore, any discussion of empirical foundations must take into account a remarkable series of papers, initiated by Kahneman and Tversky (1974), that explores the limits of rational choice theory. These papers are collected in Kahneman et al. (1982), Kahneman and Tversky (2000). The heuristics-and-biases program of Kahneman and Tversky has attracted its own critics (Gigerenzer, 1996). The critique is interesting, and has some merit. But in the end, the experimental evidence demonstrates severe limits to the power of rational choice theory (Kahneman and Tversky, 1996).

The data show that if people are trying to maximize expected utility, they don't do it very well. Errors are large and repetitive, go in predictable directions, and fall into recognizable categories. Rather than making decisions by optimization—or bounded rationality, or satisficing—people seem to use plausible heuristics that can be classified and analyzed. Rational choice theory is generally not a good basis for justifying empirical models of behaviour, because it does not describe the way real people make real choices.

Sen (2002), drawing in part on the work of Kahneman and Tversky, gives a far-reaching critique of rational choice theory, with many counter-examples to the assumptions. The theory has its place, according to Sen, but also leads to "serious descriptive and predictive problems" (p. 23). Nelson and Winter (1982) reached similar conclusions in their study of firms and industries. The axioms of orthodox economic theorizing, profit maximization and equilibrium create a "flagrant distortion of reality" (p. 21).

Almost from the beginning, there were critiques of modeling in other social sciences too. Bernert (1983) and Platt (1996) review the historical development in sociology. Abbott (1997) finds that variables like income and education are too abstract to have much explanatory power; so do models built on those variables. There is a broader examination of causal modeling in Abbott (1998). He finds that "an unthinking causalism today pervades our journals and limits our research" (p. 150). He recommends more emphasis on descriptive work and on smaller-scale theories more tightly linked to observable facts—middle-range theories, in Robert Merton's useful phrase. Clogg and Haritou (1997) consider difficulties with regression, noting that endogenous variables can all too easily be included as regressors. Hedström and Swedberg (1998) present a lively collection of essays by a number of sociologists who are quite skeptical about regression models. Rational choice theory also takes its share of criticism.

Goldthorpe (1999, 2000, 2001) describes several ideas of causation and corresponding methods of statistical proof, which have different strengths and weaknesses. He is skeptical of regression, but finds rational choice theory to be promising—unlike other scholars cited above. He favors use of descriptive statistics to infer social regularities, and statistical models that reflect generative processes. He finds the manipulationist account of causation to be generally inadequate for the social sciences. Ní Bhrolcháin (2001) has some particularly forceful examples to illustrate the limits of modeling.

Lieberson (1985) finds that in social science, non-experimental data are routinely analyzed as if they had been generated experimentally, the typical mode of analysis being a regression model with some control variables. This enterprise has "no more merit than a quest for a perpetual motion machine" (p. ix). Finer-grain analytic methods are needed for causal inference, more closely adapted to the details of the problem at hand. The role of counter-factuals is explained (pp. 45–48).

Lieberson and Lynn (2002) are equally skeptical about mimicking experimental control through complex statistical models: simple analysis of natural experiments would be preferable. Sobel (1998) reviews the literature on social stratification, concluding that "the usual modeling strategies are in need of serious change" (p. 345), also see Sobel (2000). In agreement with Lieberson, Berk (2004) doubts the possibility of inferring causation by statistical modeling, absent a strong theoretical basis for the models—which rarely is to be found.

Paul Meehl was a leading empirical psychologist. His 1954 book (Meehl, 1954) has data showing the advantage of using regression, rather than experts, to make predictions. On the other hand, his 1978 paper (Meehl, 1978), "Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology," saw hypothesis tests—and cognate black arts—as stumbling blocks that slowed the progress of psychology. Meehl and Waller (2002) discusses the choice between two similar path models, viewed as reasonable approximations to some underlying causal structure, but does not reach the critical question—how to assess the adequacy of the approximations.

Steiger (2001) provides a critical review of structural equation models. Larzalere et al. (2004) offer a more general discussion of difficulties with causal inference by purely statistical methods. Abelson (1995) has a distinctive viewpoint on statistics in psychology. There is a well-known book on the logic of causal inference, by Cook and Campbell (1979). Also see Shadish et al. (2002), who have among other things a useful discussion of manipulationist versus non-manipulationist ideas of causation.

Pilkey and Pilkey-Jarvis (2006) suggest that quantitative models in the environmental and health sciences are highly misleading. Also see Lomborg (2001), who criticizes the Malthusian position. The furor surrounding Lomborg's book makes one thing perfectly clear.

Despite the appearance of mathematical rigor and the claims to objectivity, results of environmental models are often exquisitely tuned to the sensibilities of the modelers.

In political science, after a careful review of the evidence, Green and Shapiro (1994) conclude "despite its enormous and growing prestige in the discipline, rational choice theory has yet to deliver on its promise to advance the empirical study of politics" (p. 7). Fearon (1991) discusses the role of counter-factuals. Achen (1982, 1986) provides an interesting defence of statistical models; Achen (2002) is substantially more skeptical. Dunning (2008) focuses on the assumptions behind IVLS.

King et al. (1994) are remarkably enthusiastic about regression. Brady and Collier (2004) respond with a volume of essays that compare regression methods to case studies. Invariance—together with the assumption that coefficients are constant across cases—is discussed under the rubric of causal homogeneity. The introductory chapter (Brady et al., 2004) finds that

> "it is difficult to make causal inferences from observational data, especially when research focuses on complex political processes. Behind the apparent precision of quantitative findings lie many potential problems concerning equivalence of cases, conceptualization and measurement, assumptions about the data, and choices about model specification. ... The interpretability of quantitative findings is strongly constrained by the skill with which these problems are addressed" (pp. 9–10).

There is a useful discussion in *Political Analysis* vol. 14, no. 3, summer, 2006. Also see George and Bennett (2005), Mahoney and Rueschemeyer (2003). The essay by Hall in the latter reference is especially relevant.

One of the difficulties with regression models is accounting for the ε_t's. Where do they come from, what do they mean, and why do they have the required statistical properties? Error terms are often said to represent the overall effects of factors omitted from the equation. But this characterization has problems of its own, as shown by Pratt and Schlaifer (1984, 1988).

In Holland (1986, 1988), there is a super-population model—rather than individualized error terms—to account for the randomness in causal models. However, justifying the super-population model is no easier than justifying assumptions about error terms. Stone (1993) presents a super-population model with some observed covariates and some unobserved; this paper is remarkable for its clarity.

Recently, strong claims have been made for non-linear methods that elicit the model from the data and control for unobserved confounders, with little need for substantive knowledge (Spirtes et al., 1993; Pearl, 2000). However, the track record is not encouraging (Freedman, 1997, 2004; Humphreys and Freedman, 1996, 1999). There is a free-ranging discussion of such issues in McKim and Turner (1997). Other cites to the critical literature include Oakes (1990), Diaconis (1998), Freedman (1985, 1987, 1991, 1995, 1999, 2005). Hoover (2008) is rather critical of the usual econometric models for causation, but views nonlinear methods as more promising.

Matching may sometimes be a useful alternative to modeling, but it is hardly a universal solvent. In many contexts there will be little difference between matching and modeling, especially if the matching is done on the basis of statistical models, or data from the matching are subjected to model-based adjustments. For discussion and examples, see Glazerman et al. (2003); Arceneaux et al. (2006); Wilde and Hollister (2007); Berk and Freedman (2008); *Review of Economics and Statistics*, February (2004) vol. 86, no. 1; *Journal of Econometrics*, March–April (2005) vol. 125, nos. 1–2.

3. RESPONSE SCHEDULES

The response-schedule model is the bridge between regression and causation. This model was proposed by Neyman (1923). The paper is in Polish, but there is an English translation by Dabrowska and Speed (1990) in *Statistical Science*, with discussion. Scheffé (1956) gave an expository treatment. The model was rediscovered a number of times, and was discussed in elementary textbooks of the 1960s: see Hodges and Lehmann (1964, section 9.4). The setup is often called "Rubin's model:" see for instance Holland (1986, 1988), who cites Rubin (1974). That simply mistakes the history.

Neyman's model covers observational studies—in effect, assuming these studies are experiments after suitable controls have been introduced. Indeed, Neyman does not require random assignment of treatments, assuming instead an urn model. The model is non-parametric, with a finite number of treatment levels.

Response schedules were developed further by Holland (1986, 1988) and Rubin (1974) among others, with extensions to real-valued treatment variables and parametric models, including linear causal relationships. Response schedules help clarify the process by which causation can be, under some circumstances, inferred by running regressions on observational data. The mathematical elegance of response schedules should not be permitted to obscure the basic issue. To what extent are the assumptions valid, for the applications of interest?

4. EVALUATING MODELS

One chapter in *Statistical Models: Theory and Practice* discussed a regression model for McCarthyism (Gibson, 1988). Other chapters considered a probit model for the effect of Catholic schools (Evans and Schwab, 1995), a simultaneous-equation model for education and fertility (Rindfuss et al., 1980), and a linear probability model for social capital (Schneider et al., 1997). In each case, there were serious difficulties. The studies were at the high end of the social science literature. They were chosen for their strengths, not their weaknesses. The problems are not in the studies, but in the modeling technology. More precisely, bad things happen when the technology is applied to real problems—without validating the assumptions behind the models. Taking assumptions for granted is what makes statistical techniques into philosophers' stones.

5. SUMMING UP

In the social and behavioral sciences, far-reaching claims are often made for the superiority of advanced quantitative methods—by those who manage to ignore the far-reaching assumptions behind the models. In section 2, we saw there was considerable skepticism about disentangling causal processes by statistical modeling. Freedman (2005) examined several well-known modeling exercises, and discovered good reasons for skepticism. Some kinds of problems may yield to sophisticated statistical technique; others will not. The goal of empirical research is—or should be—to increase our understanding of the phenomena, rather than displaying our mastery of technique.

REFERENCES

Abbott, A. (1997). Of time and space: The contemporary relevance of the Chicago school. *Social Forces*, 75, 1149–82.

Abbott, A. (1998). The causal devolution. *Sociological Methods and Research*, 27, 148–81.

Abelson, R.P. (1995). *Statistics as Principled Argument*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Achen, C.H. (1982). *Interpreting and Using Regression*. Sage Publications.

Achen, C.H. (1986). *The Statistical Analysis of Quasi-Experiments*. Berkeley: University of California Press.

Achen, C.H. (2002). Toward a new political methodology: Microfoundations and ART. *Annual Review of Political Science*, 5, 423–50.

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

Arceneaux, K., A.S. Gerber and D.P. Green (2006). Comparing experimental and matching methods using a large-scale voter mobilization experiment. *Political Analysis*, 14, 37–62.

Beck, N. (2001). Time-series cross-section data: What have we learned in the past few years? *Annual Review of Political Science*, 4, 271–93.

Berk, R.A. (2004). *Regression Analysis: A Constructive Critique*. Sage Publications.

Berk, R.A. and D.A. Freedman (2008). On weighting regressions by propensity scores. *Evaluation Review*, 32, 392–409.

Bernert, C. (1983). The career of causal analysis in American sociology. *British Journal of Sociology*, 34, 230–54.

Brady, H.E., D. Collier and J. Seawright (2004). Refocusing the discussion of methodology. In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, ed. H.E. Brady and D. Collier. Lanham, MD: Rowman & Littlefield Publishers, Inc, 3–20.

Brady, H.E. and D. Collier (2004). *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield Publishers, Inc, 3–20.

Clogg, C.C. and A. Haritou (1997). The regression method of causal inference and a dilemma confronting this method. In *Causality in Crisis?* ed. V.R. McKim and S.P. Turner, University of Notre Dame Press, 83–112.

Cook, T.D. and D.T. Campbell (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Chicago: Rand McNally.

Dewald, W.G., J.G. Thursby and R.G. Anderson (1986). Replication in empirical economics: The Journal of Money, Credit and Banking Project. *American Economic Review*, 76, 587–603.

Diaconis, P. (1998). A place for philosophy? The rise of modeling in statistics. *Quarterly Journal of Applied Mathematics*, 56, 797–805.

Diaconis, P. and D.A. Freedman (1998). Consistency of Bayes estimates for nonparametric regression: Normal theory. *Bernoulli Journal*, 4, 411–44.

Eaton, M.L. and D.A. Freedman (2004). Dutch book against some "objective" priors. *Bernoulli Journal*, 10, 861–72.

Engle, R.F., D. F. Hendry and J. F. Richard (1983). Exogeneity. *Econometrica*, 51, 277–304.

Evans, W.N. and R.M. Schwab (1995). Finishing high school and starting college: Do Catholic schools make a difference? *Quarterly Journal of Economics*, 110, 941–74.

Fearon, J. (1991). Counterfactuals and hypothesis testing in political science. *World Politics*, 43, 169–95.

Freedman, D.A. (1985). Statistics and the scientific method. In *Cohort Analysis in Social Research: Beyond the Identification Problem*, ed. W.M. Mason and S.E. Fienberg (with discussion). Springer, 343–90.

Freedman, D.A. (1987). As others see us: A case study in path analysis. *Journal of Educational Statistics*, 12, 101–223. Reprinted in *The Role of Models in Nonexperimental Social Science*, ed. J. Shaffer (1992), Washington, D.C.: American Educational Research Association and American Statistical Association, 3–125.

Freedman, D.A. (1991). Statistical models and shoe leather. In *Sociological Methodology*, ed. P. Marsden, chapter 10 (with discussion). Washington, D.C.: American Sociological Association.

Freedman, D.A. (1995). Some issues in the foundation of statistics. *Foundations of Science*, 1, 19–83 (with discussion). Reprinted in *Topics in the Foundation of Statistics*, ed. B.C. van Fraassen (1997). Dordrecht: Kluwer.

Freedman, D.A. (1997). From association to causation via regression. In *Causality in Crisis?* ed. V.R. McKim and S.P. Turner (with discussion). University of Notre Dame Press, 113–82. Reprinted in (1997) *Advances in Applied Mathematics*, 18, 59–110.

Freedman, D.A. (1999). From association to causation: Some remarks on the history of statistics. *Statistical Science*, 14, 243–58. Reprinted in (1999) *Journal de la Société Française de Statistique*, 140, 5–32 and in (2003) *Stochastic Musings: Perspectives from the Pioneers of the Late 20th Century*, ed. J. Panaretos. Hillsdale, NJ: Lawrence Erlbaum Associates, 45–71.

Freedman, D.A. (2004). Graphical models for causation, and the identification problem. *Evaluation Review*, 28, 267–93. Reprinted in (2005) *Identification and Inference for*

Econometric Models: Essays in Honor of Thomas Rothenberg, ed. D.W.K. Andrews and J.H. Stock, Cambridge University Press.

Freedman, D.A. (2005). *Statistical Models: Theory and Practice*. Cambridge University Press.

Freedman, D.A. (2008). Diagnostics cannot have much power against general alternatives. Forthcoming in *Journal of Forecasting*. http://www.stat.berkeley.edu/users/census /nopower.pdf.

George, A.L. and A. Bennett (2005). *Case Studies and Theory Development in the Social Sciences*. MIT Press.

Gibson, J.L. (1988). Political intolerance and political repression during the McCarthy red scare. Heinz Eulau Award from the American Political Science Association, as best paper published in 1988 in the American Political Science Review. *American Political Science Review*, 82, 511–29.

Gigerenzer, G. (1996). On narrow norms and vague heuristics. *Psychological Review*, 103, 592–96.

Glazerman, S., D.M. Levy and D. Myers (2003). Nonexperimental versus experimental estimates of earnings impacts. *Annals of the American Academy of Political and Social Science*, 589, 63–93.

Goldthorpe, J.H. (1999). *Causation, Statistics and Sociology*. Twenty-ninth Geary Lecture, Nuffield College, Oxford. The Economic and Social Research Institute: Dublin, Ireland.

Goldthorpe, J.H. (2000). *On Sociology: Numbers, Narratives, and Integration of Research and Theory*. Oxford University Press.

Goldthorpe, J.H. (2001). Causation, statistics, and sociology. *European Sociological Review*, 17, 1–20.

Green, D.P. and I. Shapiro (1994). *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science*. Yale University Press.

Heckman, J.J. (2000). Causal parameters and policy analysis in economics: A twentieth century retrospective. *The Quarterly Journal of Economics*, 115, 45–97.

Hedström, P. and R. Swedberg (1998). *Social Mechanisms*. Cambridge University Press.

Hendry, D.F. (1980). Econometrics—Alchemy or Science? *Economica*, 47, 387–406. Reprinted in D.F. Hendry (2000) chapter 1. Oxford: Blackwell.

Hodges, J.L.Jr. and E. Lehmann (1964). *Basic Concepts of Probability and Statistics*. Holden-Day: San Francisco. 2nd ed. reprinted by (2005) SIAM: Philadelphia.

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 8, 945–70.

Holland, P.W. (1988). Causal inference, path analysis, and recursive structural equation models. In *Sociological Methodology*, ed. C. Clogg (1988), chapter 13. Washington, D.C.: American Sociological Association.

Hoover, K.D. (2008). Causality in economics and econometrics. In *The New Palgrave Dictionary of Economics*, ed. S. Durlauf and L.E. Blume. 2nd ed. Macmillan.

Hubbard, R., D.E. Vetter and E.L. Little (1998). Replication in strategic management: Scientific testing for validity, generalizability, and usefulness. *Strategic Management Journal*, 19, 243–54.

Kahneman, D., P. Slovic and A. Tversky (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Kahneman, D. and A. Tversky (1974). Judgment under uncertainty: Heuristics and bias. *Science*, 185, 1124–31.

Kahneman, D. and A. Tversky (1996). On the reality of cognitive illusions. *Psychological Review*, 103, 582–91.

Kahneman, D. and A. Tversky (2000). *Choices, Values, and Frames*. Cambridge University Press.

King, G., R.O. Keohane and S. Verba (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton University Press.

Keynes, J.M. (1939). Professor Tinbergen's method. *The Economic Journal*, 49, 558–68.

Keynes, J.M. (1940). Comment [on Tinbergen's reply]. *The Economic Journal*, 50, 154–56.

Larzalere, R.E., B.R. Kuhn and B. Johnson (2004). The intervention selection bias: An underrecognized confound in intervention research. *Psychological Bulletin*, 130, 289–303.

Leamer, E.E. (1978). *Specification Searches*. Wiley.

Lieberson, S. (1985). *Making it Count*. Berkeley: University of California Press.

Lieberson, S. and F.B. Lynn (2002). Barking up the wrong branch: Alternatives to the current model of sociological science. *Annual Review of Sociology*, 28, 1–19.

Liu, T.C. (1960). Under-identification, structural estimation, and forecasting. *Econometrica*, 28, 855–65.

Lomborg, B. (2001). *The Skeptical Environmentalist: Measuring the Real State of the World*. Cambridge University Press.

Lucas, R.E.Jr. (1976). Econometric policy evaluation: A critique. In *The Phillips Curve and Labor Markets*, ed. K. Brunner and A. Meltzer. The Carnegie-Rochester Conferences on

Public Policy, supplementary series to the Journal of Monetary Economics (with discussion). Amsterdam: North-Holland, 1, 19–64.

Mahoney, J. and D. Rueschemeyer (2003). *Comparative Historical Analysis in the Social Sciences*. Cambridge University Press.

Manski, C.F. (1995). *Identification Problems in the Social Sciences*. Harvard University Press.

McKim, V.R. and S.P. Turner (1997). *Causality in Crisis?* Proceedings of the Notre Dame Conference on Causality. University of Notre Dame Press.

Meehl, P.E. (1954). *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis: University of Minnesota Press.

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806–34.

Meehl, P.E. and N.G. Waller (2002). The path analysis controversy: A new statistical approach to strong appraisal of verisimilitude. *Psychological Methods* (with discussion), 7, 283–337.

Nelson, R.R. and S.G. Winter (1982). *An Evolutionary Theory of Economic Change*. Harvard University Press.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* (in Polish), 10, 1–51. English translation by D.M. Dabrowska and T.P. Speed (1990). *Statistical Science* (with discussion), 5, 465–80.

Ní Bhrolcháin, M. (2001). “Divorce effects” and causality in the social sciences. *European Sociological Review*, 17, 33–57.

Oakes, M.W. (1990). *Statistical Inference*. Chestnut Hill, MA: Epidemiology Resources, Inc.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pilkey, O.H. and L. Pilkey-Jarvis (2006). *Useless Arithmetic*. Columbia University Press.

Platt, J. (1996). *A History of Sociological Research Methods in America*. Cambridge University Press.

Pratt, J.W. and R. Schlaifer (1984). On the nature and discovery of structure. *Journal of the American Statistical Association* (with discussion), 79, 9–33.

Pratt, J.W. and R. Schlaifer (1988). On the interpretation and observation of laws. *Journal of Econometrics*, 39, 23–52.

Rindfuss, R.R., L. Bumpass and C.St. John (1980). Education and fertility: Implications for the roles women occupy. *American Sociological Review*, 45, 431–47.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.

Scheffé, H. (1956). Alternative models for the analysis of variance. *Annals of Mathematical Statistics*, 27, 251–71.

Schneider, M., P. Teske and M. Marschall (1997). Institutional arrangements and the creation of social capital: The effects of public school choice. *American Political Science Review*, 91, 82–93.

Sen, A.K. (2002). *Rationality and Freedom*. Harvard University Press.

Shadish, W.R., T.D. Cook and D.T. Campbell (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Sims, C.A. (1980). Macroeconomics and reality. *Econometrica*, 48, 1–47.

Sobel, M.E. (1998). Causal Inference in Statistical Models of the Process of Socioeconomic Achievement: A Case Study. *Sociological Methods Research*, November 1, 27(2), 318–348.

Sobel, M.E. (2000). Causal inference in the social sciences. *Journal of the American Statistical Association*, 95, 647–51.

Spirtes, P., C. Glymour and R. Scheines (1993). Causation, Prediction, and Search. *Springer Lecture Notes in Statistics*, 81. 2nd ed. (2000). MIT Press.

Steiger, J.H. (2001). Driving fast in reverse. *Journal of the American Statistical Association*, 96, 331–38.

Stone, R. (1993). The assumptions on which causal inferences rest. *Journal of the Royal Statistical Society, Series B*, 55, 455–66.

Tinbergen, J. (1940). On a method of statistical business-cycle research. A reply [to Keynes]. *The Economic Journal*, 50, 141–54.

Wilde, E.T. and R. Hollister (2007). How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment. *Journal of Policy Analysis and Management*, 26, 455–77.