Causal Inference and Evidence-based Policy
ECON 8899; PMAP 8899; CRJU 8899
Time: M-W 1:30 pm – 2:45 pm
Location: 631 Langdale

Professor: Paul J. Ferraro
Office: 458 14 Marietta Street (4th Floor, Andrew Young Policy School Building)
Phone: 404-413-0201
e mail: pferraro@gsu.edu
Course Web: Desire2Learn

Office Hours: Wednesday 12:15 pm to 1:15 pm and by appointment.

Course Description
Did California's Proposition 99 reduce sales of tobacco products in the state? Did the Clean Air Act result in cleaner air? Do charter schools increase student achievement and, if so, what types of students respond most? Cause and effect questions like these motivate much of the empirical work in policy sciences. Does X cause Y? If X causes Y, does it cause Y in all situations? Through what mechanisms does X cause Y? If X causes Y, how large is the effect of X on Y and how does the size compare to other causes of Y? To answer these cause-and-effect questions, a counterfactual model of causality and a unified methodological framework has been developed.

This class aims to teach students to apply and interpret the counterfactual model and associated methods in answering policy-relevant empirical questions. The course has a heavy reading load with an emphasis on readings that elucidate the intuition and the application of the core conceptual ideas. I am a firm believer that the most fundamental principles can be stated in plain English. Thus the course stresses intuition (in English) over mechanics and proofs. Nevertheless students will be expected to apply the mechanics in homework and in a term paper.

Whether you are a student with substantial graduate work in empirical methods or a student with only the pre-requisites covered, you should expect to gain a deeper understanding of approaches to answering causal questions and of the nature of evidence itself. Importantly, you will see more clearly the connections among the various approaches to estimating causal effects. Even for students with substantial coursework in statistics, these connections are often missed.

I have outlined a provisional syllabus below, but we can adapt it based on student interests and background. The main emphasis of the course is like any other graduate course: to encourage students to think critically, to speak and write simply and clearly, to own and use a body of facts and ideas that are widely known, to detect errors and fallacies, to resolve intellectual problems, and to advance our collective knowledge through independent research.

Course Prerequisites
A graduate-level statistics course that covers probability theory, hypothesis testing and linear regression (examples include ECON8740 and ECON8840; ECON 9710 and 9720; PMAP 9111 and PMAP 9121; CRJU 8620 and CRJU 9630). Contact the professor if you are unsure whether your background is sufficient for the course.

Required Textbooks

Morgan, SL and C Winship. 2007. Counterfactuals and Causal Inference: methods and principles for social research.

Angrist, JD and JS Pischke. 2008. Mostly Harmless Econometrics: an empiricist's companion. Princeton University Press.

The course reading list will also include journal articles. Most of these are available via the Pullen Library electronic journals portal (http://wwwlib.gsu.edu/ejournals/). Articles that are not available via Georgia State University Library will be posted on Desire2Learn.

Recommended Textbooks

Rosenbaum, P. 2010. *Observational Studies*. Springer. Given this book's expense, it is not required, but I do recommend it if you plan to do empirical work for a career. Copies of a few relevant chapters will be available online. If you plan to do field experiments in your research, I recommend Gerber and Green's *Field Experiments: design, analysis and interpretation* (Larry Orr's *Social Experiments* text is dated, but filled with useful practical information).

Grading

I will grade you based on your performance on homework problems, class participation (including keeping up with the readings), and a term paper. Please remember that all university regulations, deadlines, and policies must be observed (incl. the Policy on Academic Honesty).

Homework Problems (35% of grade)

The problem sets are designed to bridge the gap between the teaching of theory and the application of that theory. By doing the homework sets, students will make the theory operational and will apply statistical software to do causal analysis. You can choose to use SPSS, STATA or R software (other programs only with permission of instructor). Students can work together to solve the problems, but each student must hand in his or her own work.

Final Paper (60% of grade)

See p.12 of the syllabus. The final paper is designed to encourage you to do original research on a policy topic of your choice. You are strongly encouraged to choose your topic from the professor's list of suggested topics. You need special permission to choose a topic that is not on the list. We will discuss the paper requirements in more detail in class. The basic idea of the paper is that it should be publishable *somewhere* (whether you publish it or not is your choice). I am not asking for a major breakthrough that is destined for a top journal. <u>Choose something manageable</u> that demonstrates you can recognize an interesting causal question, can apply your knowledge of evaluation design, statistical methods and the real world to attempt to answer the question, and can communicate your results clearly, concisely and cogently. A well-posed and answered "small" question is much more desirable than an ambitious, but convoluted and opaque tome. Students tend to do best when they take a published empirical article for which they can obtain the data set, or take one of the professor's available data sets, and then replicate the original analysis and extend it in some small but useful way (e.g., update data, apply different methods). The <u>approximate</u> weights are:

(25%) Explanation of the causal relationship of interest, the ideal experiment, and the identification strategy.
(55%) Analysis and interpretations of results
(15%) Caveats, implications, and suggestions for future research.
(5%) 10 minute presentation of your draft paper to the class (if >24 students in class, we may eliminate this presentation and put weight into Analysis and Interpretations).

The paper topic must be selected in consultation with me. Start thinking about it early in the semester. A **two-paged proposal** must be handed in by **6 October**. The earlier you hand in a proposal, the more input I can have in directing your research. A **3-5 page proposal** with detailed empirical design and methods description and, if necessary, a more clearly formulated research question is due on **15 October**. Each of you will send me an email the week of 10 November to describe progress of your paper and any problems you may be having.

I advise you to complete your preliminary results by 19 November and generate a draft paper by 1 December (not to hand in, just a recommended deadline for yourself). Class presentations will begin on 4 December and will continue, if necessary, through the final exam period. The final paper is due at 5 pm on the day of the regularly scheduled final exam (10 December).

Class Participation (5% of grade)

Class participation essentially means (1) that you show up for most classes (no need to give me excuses for missed classes) and (2) you show up having read the assigned readings on most days. I will speak more about what I mean by class participation on the first day of class. As a commitment device, you will often be asked to type up a half page of comments and questions on each reading and bring it to class. These submissions will count towards class participation. If the class is small enough, I may also assign students to do a brief presentation on one of the readings. There is no better way to learn a topic than to try to teach it to others.

Preliminary Course Outline

I. Causal States, Potential Outcomes, Identification and Treatment Effects
II. Experimental Designs
Selection on Observables
III. Estimating Causal Effects by Conditioning: Matching
IV. Estimating Causal Effects by Conditioning: Regression
Selection on Unobservables
V. Partial Identification, Sensitivity Analyses, Multiple Control Groups, and Tests of Known Effects
VI. Instrumental Variable Designs of Causal Effects
VII. Regression Discontinuity Designs
VIII. Explanation and Identification of Causal Effects by Mechanisms
IX. Repeated Observations (Panel Data) and Estimation of Causal Effects
Other topics
X. Heterogeneous Treatment Effects
XI. Synthetic Controls and Comparative Case Studies
XII. What Constitutes "Evidence?"

Reading Assignments
In general, we will use the required textbooks for the key concepts (supplemented with other texts and review articles, when needed). We will use journal articles as examples of applications. Starting in September, I will attempt to achieve the following structure: (1) assign chapters on theory and at least one applied article; and (2) the next week will "look back" at the previous week with a second applied article and then introduce new theory. Below is a draft list of the reading assignments. I expect I will alter the list as we go, depending on student interest and revealed abilities. Journal articles are available through GSU library journal locator portal. Readings that are not available through this portal will be put on Desire2Learn.

*** indicates required reading.** Readings without the asterisks will be described in class, but you need not read them in depth (but do skim them). If other readings are assigned, you will receive them well in advance.

An overview of the class topics in one article can be found in Ferraro and Hanauer (In Press):
http://www2.gsu.edu/~wwwcec/docs/ARER%20PrePrint%20Ferraro%20Hanauer.pdf

I. Causal States, Potential Outcomes, Identification, Treatment Effects and Elaborate Theories

Morgan and Winship, Chapters 1-2*
Angrist and Pischke, Chapters 1-2*
Rosenbaum, Chapter 1* [Desire2Learn]

Morgan and Winship, Chapter 10
Rosenbaum, Chapter 11.6

[If you've never heard of any of these ideas before or would appreciate more basic articles beyond the Ferraro and Hanauer article above, try: Ravallion, M. 2001. The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluation. The World Bank Economic Review 15(1): 115-140////and////Ferraro, PJ. 2009. Counterfactual thinking and impact evaluation in environmental policy. In Special Issue on Environmental program and policy evaluation, M. Birnbaum & P. Mickwitz (Eds.). *New Directions for Evaluation* 122: 75–84.]

For more technical discussions of causality, see Holland (1986; *JASA*), Heckman (2000; *QJE*), and, especially, J. Pearl's book *Causality*.

II. Experimental Designs

Duflo et al. 2006. Using randomization in development economics research: a toolkit.*

Gerber and Green. 2012. Field Experiments. Chapter 2* [Desire2Learn] Chapter 12 recommended.

Rosenbaum, Chapter 2 (through section 2.4) * [Desire2Learn]

Ferraro, PJ, M Price. 2013. Using Non-pecuniary Strategies to Influence Behavior: evidence from a large-scale field experiment. *The Review of Economics and Statistics* 95(1): 64-73.* {homework is based on this article, but we will not discuss much in class}

Thornton, R. 2008. The Demand for, and Impact of, Learning HIV Status. American Economic Review 98(5): 1829-1863.*

Bloom, H.S., L.L. Orr, S. H. Bell, G. Cave, F. Doolittle, W. Lin and J.M. Bos. 1997. The Benefits and Costs of the JTPA Title II-A Programs. *Journal of Human Resources*, 32: 549-576.

Miguel, E and M Kremer. 2004. Worms: Education and Health Externalities in Kenya. *Econometrica* 72(1)

Olken. 2007. Monitoring Corruption: evidence from a field experiment in Indonesia. *Journal of Political Economy*.

Ludwig, Jens, Duncan, Greg, and Hischfield, Urban Poverty and Juvenile Crime: Evidence from a Randomized Housing-Mobility Experiment. *Quarterly Journal of Economics*, 116(2), May 2001, 655-679.

Gerber, A and D Green. 2000. The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: a field experiment. American Political Science Review 94(3): 653-663.

Orr, L. 1999. Social experiments: evaluating public programs with experimental methods. Chapters 1, 2 and 6 are recommended for issues like ethics, informed consent, relating experimental results to policy and other practical issues.

III. Estimating Causal Effects by Conditioning: Matching

Morgan and Winship, Chapters 3-4*

Ho D, Imai K, King G, Stuart E (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit Anal* 15**,** 199-236.*

Andam, KS, PJ Ferraro, A Pfaff, GA Sanchez-Azofeifa, and J Robalino. 2008. Measuring the Effectiveness of Protected Area Networks in Reducing Deforestation. *Proceedings of the National Academy of Sciences* 105(42): 16089-16094.* {homework is based on this article, but we will not discuss much in class}

Smith, J and P Todd. 2005. Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators. *Journal of Econometrics*.*

Shadish et al. 2008. Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments. *JASA* 103(484).

[for more technical material on matching, see Rosenbaum, Chapter 3]

IV. Estimating Causal Effects by Conditioning: Regression

Morgan and Winship, Chapter 5.*
Angrist and Pischke, Chapter 3.*

Bitler MP, Currie J. 2005. Does WIC work? The effects of WIC on pregnancy and birth outcomes. Journal of Policy Analysis and Management. 24(1): 73-91.

Crump et al. 2009. Dealing with Limited Overlap in Estimation of Average Treatment Effects. Biometrika 96(1).

V. Partial Identification, Sensitivity Analyses and Tests of Known Effects

Morgan and Winship. Chapter 6*

Rosenbaum. Chapter 4* [Desire2Learn]

Manski, C. and D. Nagin (1998). Bounding Disagreements about Treatment Effects: A Case Study of Sentencing and Recidivism. Sociological Methodology, vol. 28, pp. 99-137. *
[in class exercise on bounding the effects of the death penalty]

Arriagada, PJ Ferraro, S Pattanayak, R, E Sills, and S Cordero. 2012 Do payments for environmental services reduce deforestation? A farm-level evaluation from Costa Rica. *Land Economics*.* [just read through the partial identification section]

Manksi, C. 2005. Social Choice with Partial Knowledge of Treatment Response. Chapters 1-2.

Manksi, C. 2011. Policy analysis with incredible certitude. The Economic Journal, 121 [every student thinking of doing policy research should at least skim this article]

Dinardo JE and J-S Pischke. 1997. The Returns to Computer Use Revisited: have pencils changed the wage structure too? Quarterly Journal of Economics.

Altonji et al. 2005. Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic School. Journal of Political Economy 113(1)

Rosenbaum, Chapter 6

VI. Instrumental Variable Designs of Causal Effects

Morgan and Winship, Chapter 7*
[in class exercise on assessing the quality of instrumental variable designs]

Angrist and Pischke, Chapter 4*

Heckman, J. 1997. Instrumental Variables: a study of implicit behavioral assumptions used in making program evaluations. *The Journal of Human Resources* 932(3):441-462.* {focus on his critique of the behavioral assumptions underlying IV estimators}

Hanson, A. 2009. Local employment, poverty, and property value effects of geographically-targeted tax incentives: an instrumental variable approach. Regional Science and Urban Economics 39: 721-731.* {homework is based on this article, but we will not discuss much in class}

Angrist, Joshua D. and William N. Evans. Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. *The American Economic Review* 88(3): 450-477.

Angrist, J.D., G.W. Imbens, and D. B. Rubin (1996) "Identification of Causal Effects Using Instrumental Variables," Journal of the American Statistical Association, 91: 444-472.

James Heckman. Instrumental Variables: A Study of Implicit Behavioral Assumptions. *Journal of Human Resources*.

Joshua Angrist and Guido Imbens, "Comment on 'Instrumental Variables: A Study of Implicit Behavioral Assumptions,'" *JHR*.

James Heckman, Reply to previous, *JHR*.

Walker and Currie. 2011. Traffic Congestion and Infant Health: Evidence from E-ZPass. American Economic Journal: Applied Economics 3(1): 65-90. [not an IV design, but similar idea of finding natural sources of variation in treatment assignment that is unrelated to potential treatment states and unrelated to potential outcomes except through the treatment]

VII. Regression Discontinuity Designs

Morgan and Winship, Chapter 9.2*

Angrist and Pischke, Chapter 6*

McIntosh, C. J Alix-Garcia, K Sims, and J Welch. 2013 The Ecological Footprint of Poverty Alleviation: Evidence from Mexico's Oportunidades Program. *Review of Economics and Statistics*.*

Black, DA, J Galdo, and JA Smith. Evaluating the worker profiling and reemployment services using a regression discontinuity approach. American Economic Review Papers and Proceedings, 97(2): 104-107.

Chay, KY and M Greenstone. 2003. Air Quality, Infant Mortality and the Clean Air Act of 1970. Journal of Political Economy

Boudelmeyer and Skoufias. An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA.

VIII. Repeated Observations and Estimation of Causal Effects
{because of time constraints, this lecture and reading list will necessarily be abbreviated; the key is to understand the assumptions under which common panel data estimators estimate a relevant causal effect}

Morgan and Winship, Chapter 9.1, 9.3-9.4*

Angrist and Pischke, Chapter 5* [also read Chapter 8 on "nonstandard standard errors"]

Card, D. and A. B. Krueger (1994) "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania .American Economic Review, 84: 772-793.*

Ferraro, PJ, JJ Miranda. Can panel data designs and estimators substitute for randomized controlled trials in program impact evaluations? Working paper.*

Arriagada, PJ Ferraro, S Pattanayak, R, E Sills, and S Cordero. 2012 Do payments for environmental services reduce deforestation? A farm-level evaluation from Costa Rica. *Land Economics*.

Acemoglu, D., J. Angrist. 2001. Consequences of Employment Protection? The Case of the American with Disabilities Act. *Journal of Political Economy*, 109(5).

Guyran, Jonathan. 2004. Desegregation and black dropout rates. American Economic Review 94: 919-943.

IX. *Mechanism Causal Effects and Identification of Average Causal Effects by Mechanisms*

Morgan and Winship, Chapter 8 (re-read Chapter 3 on causal graphs)*

Ferraro, PJ and M Hanauer. 2013. Causal mechanism effects of protected areas on poverty through changes in ecosystem services and infrastructure development. {available online}*

X. Heterogeneous Treatment Effects

Angrist and Pischke, Chapter 7 (quantile regressions)*

Orr, Chapter 6.* [Desire2Learn]

Ferraro, PJ and JJ Miranda Montero. 2013. Long-term, heterogeneous treatment effects from non-pecuniary environmental programs: a large-scale field experiment. Resources and Energy Economics*

XI. Synthetic Controls and Comparative Case Studies

Abadie et al. 2010. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *JASA*.*

Abadie A and J. Gardeazabal. 2003. The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*.

XII. What Constitutes Evidence?

[Exchange in JEL issue]* [Desire2Learn]

Learning Objectives

1. The student should be able to understand and apply the basic theory from the prerequisite classes.

2. The student should be able to explain what is meant by causal inference in the counterfactual framework.

3. The student, when faced with a specific policy problem, should be able to identify key attributes of the problem that are amenable to causal inference using experimental or quasi-experimental designs.

4. For a given causal question, the student should be able to describe an experimental design capable (in principle) of uncovering a causal relationship, and discuss the advantages and disadvantages of such a design from the perspective of causal inference, practical design and policy implications.

5. In the context of non-experimental policy or program implementation, the student should be able to select the most appropriate evaluation design(s) and methodology conditional on the characteristics of the available data and the policy question, implement these design and methods, and appropriately interpret the results and their potential biases.

6. The student should be able to evaluate the quality of evidence from a given empirical study: in other words, they should be able to articulate the causal relationship of interest, describe the identification strategy, describe the implicit assumptions on which this strategy rests, and characterize the quality of the inferences drawn.

NOTE: The course syllabus provides a general plan for the course; deviations may be necessary.

Manuscript Review

[I am no longer assigning this review in the class because the class has gotten too big, but I thought students might still like to read it. If you don't care, move on to p.12]

You will be asked to review a manuscript, which will be posted on Desire2Learn two weeks before the due date. A few of these manuscripts are not quite ready for submission to a journal, but you should treat them as if they were submitted. All focus on causal effect estimation.

The review comprises two parts. First, a short <u>cover letter</u> to the editor indicating whether you think the manuscript should be (1) accepted as is or with minor changes that are noted in your review; (2) revised and re-submitted along the lines recommended in your review; or (3) rejected. Explain why. This recommendation is for the editor only and should not be in your detailed comments to the author. Remember to consider the intended audience and whether this paper makes a contribution to the relevant literature.

Second, a <u>detailed review</u> for the authors, and which the editor can read and interpret your cover letter recommendation. Here's a suggested structure for your review (you can adapt as you see fit):
<u>General Impression</u>: Start with a paragraph that summarizes the author's point and your general impression of the article.
<u>Existing Literature</u>: A paragraph (or more) on this topic is needed if you feel that something in the literature was missed or misrepresented.
<u>Writing Style:</u> Don't spend much space/time editing the writing style, but if you think the paper was poorly written or well written, say so.
<u>Specific Comments (Major):</u> Here is where you put in your major accolades, critiques, and questions.
<u>Specific Comments (Minor):</u> Here is where you put in other critiques or unanswered questions that are not critical to ensuring the paper is a valuable contribution, but you believe could improve the paper. This is also the section where you could make editorial suggestions or point out typographical errors (however, there is no need to do so. You're <u>not</u> an editor. You're a reviewer.).

This class is a class in causal inference, so you must offer your opinions and analysis of the design and causal inferences drawn. These opinions and analyses are the most important part of the review. Think about confounding factors (those identified by author or those that you might identify) and how well the authors control for them. You are also encouraged, but not required, to make comments on other aspects of the manuscript.

Here are some guiding questions that may or may not be appropriate for your particular review.

1. What question(s) is the author trying to answer? What are the conclusions and are they consistent with the data or the analysis?
2. Is the paper well written and clear in its arguments?
3. If there is theory (in some cases it may be implicit rather than explicit), does it adequately motivate the empirical analysis? In other words, does the empirical design seem well

matched to the theory? If there is a formal theoretical model, is it specified appropriately? What are the implicit assumptions that may be violated in the empirical context studied?

4. What are the confounding factors and does the author's empirical design address them adequately? If not, is the potential bias something that would overturn the author's conclusion?
5. Does the author neglect anything from the past literature on this topic (this may require you doing a skim of the literature to familiarize yourself with it) or does the author misrepresent her contribution?
6. Identify strengths of the manuscript.
7. Identify errors or weaknesses of paper, if any (e.g., hard to justify assumptions, strange formulation of the model), or how the results may change with a realistic change in assumptions or addition of other factors (you may be able to show something analytically or with a numerical example, but if you cannot, simply discuss how the results may change if one were to change an assumption or a variable).
8. Ideas for extensions or future research on the topic.

The review (letters to the editor and authors) should be typed and 5-10 pages double-spaced (12 point font; 1 inch margins). Good reviews can generally be done in a half-page to the editor and two-pages single-spaced to authors (sometimes fewer), but I'm asking a bit more for this class assignment. There are two files: the cover letter to the editor, in which you put your name, and the letter to the authors in which you do NOT put your name.

The review is supposed to be double-blind so even if you figure out who the author is, **please do NOT contact the author [you will receive a zero on the assignment if you contact the author].** I will be giving your review to the authors without your name. Thus, in addition to the paper copy you submit in class, you should also submit your report to the authors to me electronically with no name on it.

I will post an example of a real referee report. I tried to strip all identifying information from it but if you worked hard enough, you could still figure out who the authors are. I'm giving you this example to help you in your development as an economist. Please be professional and do not distribute it or do something else stupid with it like post it on the web. I'll also give you a report I've received on one of my manuscripts.

Note that the sample review I post is just one review. Others I have done look a lot different and are a function of the paper and journal. They may be shorter, longer, much more positive, or much more negative. My comments and style reflect one professor's opinion on doing a review. So take my advice in this domain with that in mind. I will also post a couple of other guideline documents, written by others, for doing scholarly referee reports.

Lastly: Do Not Lard Your Review with BS – This is Not a Book Report

Term Paper Instructions

Read the syllabus for motivation and due dates, but use this document to guide you in writing. The paper must <u>not</u> be more than 10 pages double-spaced (12 point font) with one-inch margins on all sides, <u>excluding</u> cover page, references, figures, tables and computer code (<u>you must submit your code</u>). I will not read more than this limit. FYI…American Economic Review Papers and Proceedings are approximately 10 pages for everything, so you can write a good paper in 10 pages. Being able to write succinctly is a good skill.

Paper Elements

- Cover Page: Your name, title and 50-word abstract. Does not count towards page limit.

- Page 1: <u>Concise</u> summary of the key aspects of the paper, represented by answers to the following 5 questions. This gives me a quick guide to what I'm going to see in the paper. This page can be single-spaced & does not need references (other pages double-spaced).

1. What is the causal relationship of interest?
   a. What is the treatment?
   b. What is the outcome?
   c. What is the relevant target population?
   d. What is the theory that connects treatment to outcome in this population? Be succinct.
2. What is the treatment effect of interest?
3. What is the experiment that could ideally be used to estimate this treatment effect?
4. What is your identification strategy and what are its implicit causal assumptions? You may have two strategies (the strategy in the original paper you are replicating and an alternative one that you are doing as an extension of the original paper).
5. What is your method of statistical inference?

- Main Paper, pages 2-10

1. Describe original study (1 paragraph of ≤ 6 sentences: motivation, design, results).
2. Describe data (1 paragraph of ≤ 6 sentences).
3. Describe your extension (1 paragraph of ≤ 6 sentences)
4. Describe the identification strategy and methods in more detail (1 page maximum).
5. Results
6. Discussion – no need to summarize key findings again (the abstract does that)
   a. Comparison and Caveats. For papers doing a replication and extension, if your results differ from the original study, why? If your results do not differ, why? For all papers, what problems remain? Please focus on hidden bias, but additional issues may also arise – ≤ 3 paragraphs.
   b. What would you do next to improve your estimate of the causal relationship of interest? ≤ 2 paragraphs.

The key to writing a good paper is to demonstrate a deep understanding of what you are doing, not to show you can do technically-advanced methods. Do a careful, appropriate analysis and think deeply about identification issues. At the end, describe remaining issues that you cannot address and indicate how you might address them or did address them (e.g., you did sensitivity analysis, tests of known effects). Although it would be nice to teach the professor something new, a better strategy is to just try to do a solid performance using well-known designs and methods.

Viable Options:

1. Take a well-designed, highly-cited empirical article in a top economics or policy journal (or a top journal in your field), and (a) replicate the main causal effect estimates and (b) extend the analysis in some way (e.g., extend data in time or covariates; use a new estimation strategy; do some tests for hidden bias, like a placebo test, or try partial identification approach). The data must be publicly available, either on a website, through a colleague, or through direct request from the author. The data have to be "ready to go." I do not want people spending their time this semester creating and cleaning a data set.
2. Take one of the professor's topics and data sets (see below).
3. Take a poorly-done (you must confirm this assertion with professor), but highly-cited empirical publication in a top scientific journal, replicate the results and extend them with a more plausible identification strategy. The data must be publicly available as described in option #1.
4. Work on your ALREADY-IN-PROCESS thesis essay that involves causal inference. Professor permission will be required in September.
5. (IN SPECIAL CASES ONLY) A systematic review of a particular class of programs or policies.

Examples with Experimental Data [much harder than they look]

1. You replicate the causal effect estimate (the main treatment effect estimated if there is more than one estimated) and then analyze new outcome data (i.e., the effect of the treatment on a different outcome) or re-analyze, using appropriate methods, the original data from an experiment that suffers from some kind of problem (e.g., attrition, non-compliance, interference).
2. You have covariate data and you replicate the main original experimental results, followed by estimating heterogeneous treatment effects or mechanism treatment effects.
3. Master students: I have seen a lot of experimentalists, particularly in development economics, design multi-treatment experiments and then evaluate the p-values to determine if one treatment is more effective than another (e.g., Treatment 1 has a p-value < 0.05 and Treatment 2 does not, and the author concludes that Treatment 1 is effective and Treatment 2 is not effective, but the authors do not actually test whether the two estimated effects are statistically different (in most cases I've seen, they look like they might not be)). Review the literature and quantify how many articles do this and explain why it's a problem (or perhaps not a problem, if that's your conclusion).
4. Master students: I have seen experimental designs that fail to reject the null of ATE=0, but then the authors search for subgroup effects are report them as causal effects. Is this common? If so, quantify the proportion of articles that do so and explain why it's a problem (i.e., data-mining makes it much more likely to find an effect just by chance).

Examples with Non-Experimental (Observational) Data

1. Do a design-replication study. At the end of this document, I include a bibliography of such studies. Many of them have publicly available data on the web (e.g., LaLonde's NSW data set), but if they do not, it may be easy to obtain the data from the authors of a design-replication study. Ferraro has one of these data sets available for the Ferraro and Price water experiment. You would assess the ability of an appropriate (given the

context), non-experimental estimator, which was not used in previous studies, to replicate the experimental benchmark estimate. Then discuss why or why not your results are different.

2. Take a published observational study that used regression, matching methods or panel data methods, replicate the main result and then try an alternative matching estimator or panel data estimator <u>that you can clearly argue would be weakly superior to the original estimator (i.e., no reason to think it would be worse)</u> - don't just pick an estimator at random. If the alternative estimator is simply "no less appropriate that the original estimator", you are essentially testing robustness of the original study's results. Most of you won't have access to a new instrumental variable that the original authors did not consider, but if you do, you are welcomed to try that as well.
3. Take any observational study that uses panel data, replicate the main result and extend the analysis by pre-processing the data with matching methods and then re-estimating the original panel data model.
4. Take any observational study that uses spatial regression models, replicate the main result and extend by pre-processing the data with matching methods and then re-estimating the original model.
5. Take any observational study, replicate the main result, and add more recent outcome data or a key confounding variable that was missing in the original study.
6. Take a published observational study, replicate the results and extend in a way that tests robustness of causal inference and permits you to demonstrate your conceptual understanding of causal inference. Examples include partial identification and other approaches that use minimal assumptions to bound treatment effects (e.g., IVs without monotonicity assumption), sensitivity tests to hidden bias, tests of known effects (placebo tests), and multiple control groups (i.e., adding another control group that helps you test hypotheses about specific forms of potential bias). For example, a working paper examines the causal effect of going to July 4th parades on political party identification (Yanagizawa-Drott and Madestam). You could test how sensitive this result is to hidden bias or, better yet, you could do a test of known effects by, for example, showing that attending July 4th parades causes an outcome that everyone knows could not be caused by these parades, but could be caused by a common hidden factor, or testing the plausibility of the exclusion restriction indirectly by replicating the study with another parade day that has no political content (e.g., St Patrick's Day Parade) and could not cause people to identify with a particular political party. Any study that uses weather as an IV might be similarly appropriate for such an extension.

Ferraro's Data

1. Based on data from Ferraro and Price (2013, Review of Economics and Statistics). I have a design-replication data set that contains the Cobb County treated households and neighboring Fulton County households during the same period. Includes a panel that goes back to May 2006 for both counties, and tax assessor data on the homes (e.g., fair market value, age of home, whether the occupant is an owner) and census block data. We also have address information (finer geo-coding than census block), but accessing those data will require you to talk to the professor. Possible extensions: (a) take Treatment 2 group and the Fulton County controls and attempt to estimate the treatment effect using non-experimental methods; (b) take Treatment 3 group and try state-of-the-

art panel data methods you have learned in another class; (c) try a geographic regression discontinuity design along the county border (such designs are highly controversial, but doing it might be informative); and (d) extend the design-replication study we have already conducted by conducting a test of known effects. There are many more possibilities. We also have voting data on members of the households and you could work on measuring voting behavior differently from the way in which we measured it and try to examine heterogeneous treatment effects of the water messages as a function of past voting behavior. We might also obtain 2010-2013 this semester, so check on this if it interests you (to examine long-term effects of treatments).

2. Andam et al. PNAS 2008 data. Possible extension: If you're interested in spatial statistics, you can explore the implications of failing to address spatial issues in the analysis. I have some ideas.

FYI, R is free and students can purchase Stata using the Stata Gradplan:
http://www.stata.com/order/new/edu/gradplan.html

A Note on Requesting Someone Else's Data

Many well-known and controversial studies have publicly available (or easy-to-obtain) data sets. For example, the controversial Card and Krueger minimum wage study (AER, 1994 and follow-up studies), the controversial Pitt and Khandker microfinance study (JPE, 1998), and important social experiments like the NSW, JTPA and Moving to Opportunity. If the data set you wish to work with is not publicly available *in a clean format*, you will have to contact the authors and request their data (as noted above, having to reconstruct the data from publicly available, but "unprocessed," data sources is not acceptable). Encouraging people to send you data is not easy. People are naturally defensive when someone asks for their data because they fear someone will find a mistake in their analysis and they will look foolish. So to increase the probability of obtaining the data, you need to look as harmless as possible. Here are some suggestions for doing so (none have been verified through rigorous analysis):

1. State that you are a graduate student in a class on causal inference and the paper assignment is to replicate the results in a paper that you admire. The instructor has instructed you to contact authors to obtain the data.
2. If the data were published in a journal that requires authors make their data available for purposes of replication (e.g., American Economic Review), you should indirectly remind them of this policy; e.g., "I am a graduate student taking a class on causal inference this semester. The instructor has given us a term paper assignment to replicate the results of a paper that we admire and that was published in a journal with an official policy on making data available for replication. I chose your paper because I think it's really well done and in a topic area in which I am interested."
3. Optional (but recommended): State explicitly to the authors that you have no intention of publishing any extensions of their study, but, if you did manage to find an interesting extension, you would offer the author (and any co-authors) an opportunity to be a co-author on any resulting publication.

4. Gently remind them that you have a deadline (e.g., "I really appreciate you taking the time to make these data available. I will need to submit my draft proposal for my replication by 1 October.")

The likely responses are (a) silence; (b) "I'll get back to you" and he or she never does; and (c) "I'll get back to you" and he or she does, but after the semester is over. Send one reminder after a week if you get silence. Send a reminder one week after they indicated they would get the data to you. If they did not indicate a date, send a reminder one week after the first response. To increase your odds, send a request to two sets of authors (two studies). If you are lucky enough to receive both data sets, you can offer one to an unsuccessful classmate.

IMPORTANT: Once you get the data, try to avoid bothering the author with many questions. Two or three clarifying questions on the data or their analysis are OK. More questions would be creating a burden for the authors and they'll be less likely to help any students in the future. This rule is especially important if you ask a professor at Georgia State. PLEASE do not 'reward' them for being kind enough to provide you with data by using up a lot of their time asking them follow-up questions on how they did their analysis (they then call me and complain!).

Notes on Design-Replication Studies

The literature on the performance of non-experimental designs can be classified into four different groups. Computer simulations, meta-analyses, double randomized preference trials, and design-replication studies. *Computer simulation* studies produce controlled but artificial data varying key features that might affect outcome variables. These types of studies can provide accurate results, but the data are artificial. *Meta-analyses* studies compare different <u>studies</u> that examine the same (approximately) treatment but use different samples, different randomization designs and different methodologies. These meta-analyses yield mixed evidence on the ability of non-experimental designs to replicate the results of experimental designs. They are unable to control fully for differences across studies and thus should be interpreted with caution. *Double randomized preference trials* are experiments where some subjects are randomly assigned subjects to be in a randomized experiment, in which subjects are <u>assigned</u> to one of multiple treatments, or a non-randomized experiment, in which subjects can <u>choose</u> one of the same multiple treatments. Shadish et al. (2008) present the only example of this type of experiment. *Design replication studies*, which are also known as *within-study design*s, estimate a program's impact by using a randomized control group. Then they re-estimate the impact by using one or more nonrandomized comparison groups and econometric techniques to eliminate or mitigate observable and unobservable sources of bias. Smith and Todd (2005) on your syllabus is an example. Most are using matching estimators. Examples of regression discontinuity designs include Black, Galdo and Smith, 2007; Green et al. 2009; Buddelmeyer and Skoufias, 2004; Lamadrid-Figueroa et al., 2008. For panel data, Smith and Todd (2005), Smith and Todd, 2005; Heckman et al., 1997; Heckman et al., 1998b (all using the same data set and simple difference-in-difference estimators combined with matching).

Design Replication Studies

Roberto Agodini and Mark Dynarski, Are Experiments the Only Option? A Look at Dropout Prevention Programs, Review of Economics and Statistics, 2004, 86, 180-194,

L. S. Aiken and S. G. West and D. E. Schwalm and J Carrol and S, Hsuing, Comparison of a Randomized and Quasi-Experimental Designs in a Single Outcome Evaluation: Efficacy of a University-Level Remidial Writing Program, Evaluation Review, 1998, 22, 207-244,

Arcenaux, K, AS Gerber, and DP Green. A Cautionary Note on the Use of Matching to Estimate Causal Effects: An Empirical Example Comparing Matching Estimates to an Experimental Benchmark, Sociological Methods and Research 2010, 39, 256-282.

Dan Black and Jose Galdo and Jeffrey Smith, Evaluating the Bias of the Regression Discontinuity Design Using Experimental Data, Working paper from http://http-server.carleton.ca/~jgaldo/docs/rddpaper.pdf, January, 2007,

H. S. Bloom and C. Michalapoulos and C. J. Hill, Using Experiments to Assess Nonexperimental Comparison-Group Methods for Measuring Program Effects, Learning More from Social Experiments, Russel Sage Foundation, 2005, ed H. S. Bloom, 173-235,

Bratberg, Espen and Grasdal, Astrid and Risa, Alf Erling, Evaluating Social Policy by Experimental and Nonexperimental Methods, Scandinavian Journal of Economics, 2002, 104, 147--171

Buddelmeyer, Hielke and Skoufias, Emmanuel, An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA,World Bank and Institute for the Study of Labor, 2004, IZA Discussion Paper No. 827

Thomas D. Cook and William R. Shadish and Vivian C. Wong, Three Conditions under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons, Journal of Policy Analysis and Management, 2008, 27, 724–750,

Thomas D. Cook and William R. Shadish and Vivian C. Wong, Within-Study Comparisons of Experiments and Non-Experiments: What the Findings Imply for the Validity of Different Kinds of Observational Study, Paper presented at the French Econometric Society Meeting on Program Evaluation, 2005,

Thomas D. Cook and Peter M. Steiner, Case Matching and the Reduction of Selection Bias in Quasi-Experiments: The Relative Importance of Pretest Measures of Outcome, of Unreliable Measurement, and of Mode of Data, Psychological Methods, 2010, 15, 56-68,

Rajeev Dehejia, Practical propensity score matching: a reply to Smith and Todd, Journal of Econometrics, 2005, 125, 355-364,

Rajev Dehejia and Sadek Wahba, Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs, Journal of the American Statistical Association, 1999, 94, 1053-1062,

Dehejia, Rajeev H. and Wahba, Sadek, Propensity Score-Matching Methods for Nonexperimental Causal Studies, The Review of Economics and Statistics, 2002, 84, 151-161,

Diaz, Juan Jose and Handa, Sudhanshu, An Assessment of Propensity Score Matching as a Nonexperimental Impact Estimator: Evidence from Mexico's PROGRESA Program, J. Human Resources, 2006, XLI, 319-345,

Ferraro, PJ, J Miranda. In Press. The Performance of Non-experimental Designs in the Evaluation of Environmental Policy: a design-replication study using a large-scale randomized experiment as a benchmark. *Journal of Economic Behavior and Organization.*

Thomas Fraker and Rebecca Maynard, The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs, Journal of Human Resources, 1987, 22, 194-227,

Glazerman, Steven and Levy, Dan M. and Myers, David, Nonexperimental Versus Experimental Estimates of Earnings Impacts, The ANNALS of the American Academy of Political and Social Science, 2003, 589, 63-93,

Paul Glewwe and Michael Kremer and Sylvie Moulin and Eric Zitzewitz, Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya, Journal of Development Economics, 2004, 74, 251 - 268,

Green, Donald P. and Leong, Terence Y. and Kern, Holger L. and Gerber, Alan S. and Larimer, Christopher W., Testing the Accuracy of Regression Discontinuity Analysis Using Experimental Benchmarks, Political Analysis, 2009, 17, 400-417,

Greenberg, David H. and Michalopoulos, Charles and Robin, Philip K., Do experimental and nonexperimental evaluations give different answers about the effectiveness of government-funded training programs?, Journal of Policy Analysis and Management, 2006, 25, 523--552,

Sudhanshu Handa and John A. Maluccio, Matching the Gold Standard: Comparing Experimental and Nonexperimental Evaluation Techniques for a Geographically Targeted Program, Economic Development and Cultural Change, 2010, 58, 415-447,

James J. Heckman and V. Joseph Hotz, Choosing among alternative nonexperimental methods for estimating the impact of Social programs: The case of manpower training, Journal of the American Statistical Association, 1998, 84, 862-874, James J. Heckman and Hidehiko Ichimura and Jeffrey A. Smith and PetraE. Todd, Characterizing Selection Bias Using Experimental Data, Econometrica, 1998, 66, 1017-1098,

James J. Heckman and Hidehiko Ichimura and Petra E. Todd, Matching as an econometric evaluation estimator: evidence from evaluating a job training programme, Review of Economics and Statistics, 1997, 64, 605-654,

Hill, Jennifer; Jerome Reiter and Elaine Zanutto (2004), A Comparison of Experimental and Observational Data Analyses, In: "Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives", Andrew Gelman and Xiao-Li Meng [Ed.]; New York: Wiley.

Hotz, V. Joseph and Imbens, Guido W. and Klerman, Jacob Alex,The Long-Term Gains from GAIN: A Re-Analysis of the Impacts of the California GAIN Program,National Bureau of Economic Research, 2000, NBER Working Paper, number = 8007,Cambridge MA,

Hotz, V. Joseph and Imbens, Guido W. and Mortimer, Julie H., Predicting the Efficacy of Future Training Programs Using Past Experiences, National Bureau of Economic Research, 1999, NBER Technical Working Paper, number = 238, Cambridge MA,

Robert LaLonde, Evaluating the econometric evaluation of training with experimental data, American Economic Review, 1986, 76, 604-620,

Lalonde, Robert and Rebecca Maynard (1987), How Precise are Evaluations of Employment and Training Programs – Evidence from a Field Experiment, *Evaluation Review*, Vol. 11, No. 4, pp. 428–451.

Lamadrid-Figueroa, Hector; Gustavo Angeles; Thomas Mroz; Jose Urquieta-Salomon; Bernardo Hernandez-Pado; Aurelio Cruz-Valdez; Marta Téllez-Rojo (2008), Impact of Oportunidades on Contraceptive Methods Use in Adolescent and Young Adult Women Living In Rural Areas 1997-2000, MEASURE Evaluation Working Paper Series WP-08-109.

McKenzie, David and Gibson, John and Stillman, Steven, How Important Is Selection? Experimental vs. Non-Experimental Measures of the Income Gains from Migration, Journal of the European Economic Association, 2010, 8, 913-945,

Charles Michalopoulos and Howard S. Bloom and Carolyn J. Hill, Can Propensity-Score Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs? Review of Economics and Statistics, 2004, 86,

Robert Olsen and Paul Decker,Testing Different Methods of Estimating the Impacts of Worker Profiling and Reemployment Service Systems, Mathematica Policy REsearch Inc., 2001, Washington DC,

Steffi Pohl and Peter M. Steiner and Jens Eisermann and Renate Sollener and Thomas D. Cook, Unbiased Causal Inference From an Observational Study: Results of a Within-Study Comparison, Educational Evaluation and Policy Analysis, 2009, 31, 463-479,

William R. SHADISH, M. H. CLARK, and Peter M. STEINER, Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments, Journal of the American Statistical Association, 2008, 1334-1343 (and comments that follow the article)

Jeffrey A. Smith and Petra E. Todd, Does matching overcome LaLonde's critique of nonexperimental estimators?, Journal of Econometrics, 2005, 125, 305-353,

Jeffrey A. Smith and Petra E. Todd, Rejoinder, Journal of Econometrics, 2005, 125, 365-375,

Mengli Song and Rebecca Herman, Critial Issues and Common Pitfalls in Designing and Conducting Impact Studies in Education: Lessons Learned From the What Works Clearinghouse (Phase I), Educational Evaluation and Policy Analysis, 2010, 32, 351-371, [not original design-replication study, but interesting and related]

Peter M. Steiner and Thomas D. Cook and William R. Shadish and M.H. Clark, The Importance of Covariate Selection in Controlling for Selection Bias in Observational Studies, Psychological Methods, 2010, 15, 250 - 267,

Wilde, Elizabeth and Robinson 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*, Vol. 26, No. 3, pp. 455–477.