Propensity Score Matching:

Moving Towards Causality in Education

Kenneth L. Thompson, Ph.D.

March 31, 2016

MILLSAPS COLLEGE

Today's Topics

- Causality as a priority & a responsibility
- Planning for causal inferences
- What is a propensity score?
- Matching data using propensity scores

- Where do we go next?

Why Consider Causality?

How do we provide feedback to professors so that they can improve their practice?

How do we help professors identify effectiveness?

An example: Identifying an effective reading intervention program

- School considering purchasing a reading program advertised to increase student interest in reading.
- Use for one school year in one school and compare any increase/decrease in books read during summer befc summer after with another school using traditional r instruction methods.
- Use the results to make a decision about implementi program district-wide.

Results

Average increase in books read by students ***not*** participating in reading intervention:	.06
Average increase in books read by students participating in reading intervention:	.64
Effect size of intervention program (Cohen's *d*)	.85

Cohen's d is a measure of the relationship between two means.

Smaller value = less difference in means = less effect

Larger value = larger difference in means = larger effect

But can we *really* say the reading program caused the increase in books read?

Planning for causal inference

Experimental Design

- Considered the "gold standard" for causal inference (West, 2009)
- May not be possible in practice
- Arguments against experimental design in schools (Cook, 2002), such as students aren't randomly assigned to schools
- Can rarely be undertaken in schools
- Does not describe how schools actually make decisions

Quasi-Experimental Design

- Use group comparisons
- Typically pretest/posttest non-equivalent groups design
- Groups may systematically differ from one another based on number of variables
  - Can lead to finding a difference when there isn't one; or
  - Can lead to finding less of a difference than actually exists
- Cannot be directly compared
- Do not allow causal inferences
- Leads us to think of results as "On average…"

Bad Things Can Happen When We Take An "On Average…" Perspective

What can we do to address the
"On average…" scenario?

Propensity Score Matching:

A Technique Whose Time Has Come

Why Propensity Score Matching?

- Experimental design not always possible in practice
  - Random assignment -> Causal inferences, BUT
  - Students are not randomly assigned to schools
- Quasi-Experimental design does not allow causal inferences
  - Groups may appear to look alike when they aren't really alike on the measures that matter
  - Can lead to finding a difference when there isn't one; or
  - Can lead to finding less of a difference than a
    exists
- Propensity scores help address lack of random a
  and the need for causal inferences

Do Groups Really Look Alike?

On Average, Do The Groups Look Similar?

- 5 Red/5 Black
- 4 Diamonds/4 Spades/1 Club/1 Heart
- Two straights? (5 through 9)(10 through Ace)
- 5 through 9 = Red; 10 through Ace = Black

What Is Propensity Score Matching?

- A way to remove systematic differences in groups
- A statistical technique that aims to extend causal inference into non-randomized or quasi-experimental studies (Rosenbaum & Rubin, 1983)
- Reduce differences in groups by matching participants on their likelihood of group assignment (most often logistic regression)
- After matching causal inference can be extended
- Not widely used in social science research (Hard Anthony, 2010)

What Is A Propensity Score?

- Conditional probability of assignment to a particular group given a set of variables (Rosenbaum & Rubin, 1983)
- Variables are similar to predictors in regression
- Incorporate multiple variables into a single propensity score ranging from 0 to 1
- Propensity score can be used to match participants in groups

Picking The "Best" Variables

- No limits on the number of variables that may be used estimating propensity scores
- Use variables grounded in literature

Socioeconomic Status, Race, Age, Parental Involvement, Gender, and Phonetic Awareness

(Sénéchal & LeFevre, 2002)

Propensity Scores For All Students

	Control		Intervention	
Student	Propensity Score (π)		Student	Propensity Score (π)
Brooner, Deborah	0.05		Schumacher, Gerald	0.83
Decman, John	0.45		McDonald, Denise	0.66
Brown, Sue	0.03		McEnery, Lillian	0.19
Browning, Sandra	0.15		Seevers, Randy	0.84
Jones, Robert	0.82		Orange, Amy	0.75
Carman, Carol	0.39		Peters, Michelle	0.76
Kajs, Larry	0.36		Shermis, Mark	0.63
Gavins, Marva	0.53		Simieou, Felix	0.57
Grigsby, Bettye	0.20		Vesey, Winona	0.90
Morgan, Bryan	0.50		Spuck, Dennis	0.45
Huss-Kessler, Rebecca	0.50		Stockton, Carl	0.73
Pitre, Esrom	0.56		Weaver, Laurie	0.11
Jones, Lisa	0.15		Staples, William	0.46
Kahn, Michele	0.21		Wagner, Paul	0.79
Marquez, Judith	0.32			
Matthew, Kathryn	0.08			

Matching Techniques

- Which ones to use?
  - Exact
  - Nearest Neighbor
  - Caliper
  - Nearest Neighbor within a Caliper
  - Optimal

Matching Techniques

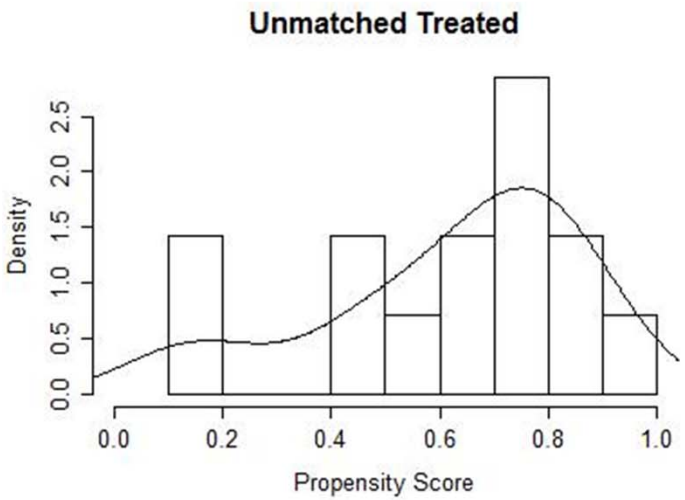
- How many to use?
  - One-to-One
  - One-to-Many

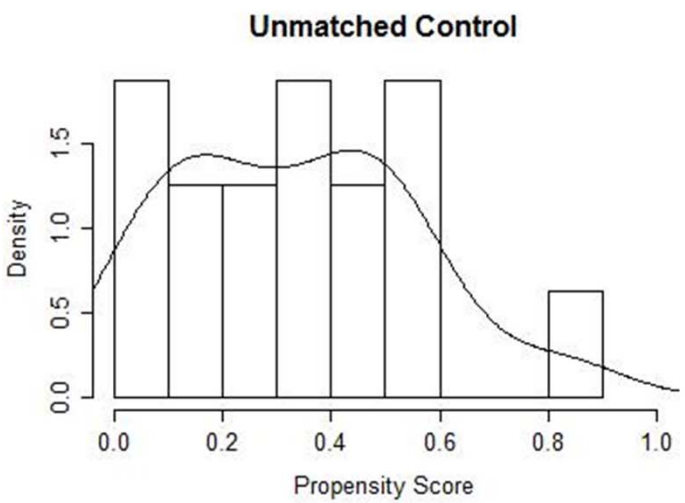
Once matched, treatment effects should be more reflective of the true effect and analogous to interpretation of randomized designs

Matching Our Groups

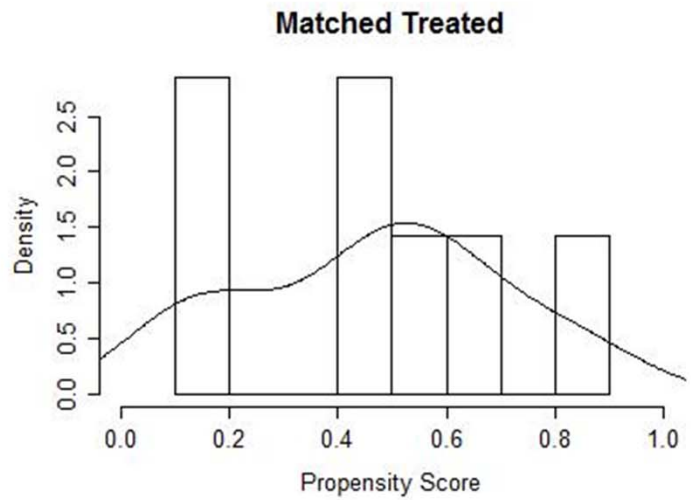
	Control		Intervention	
Student	Propensity Score (π)		Student	Propensity Score (π)
Brooner, Deborah	0.05		Schumacher, Gerald	0.83
Decman, John	0.45		McDonald, Denise	0.66
Brown, Sue	0.03		McEnery, Lillian	0.19
Browning, Sandra	0.15		Seevers, Randy	0.84
Jones, Robert	0.82		Orange, Amy	0.75
Carman, Carol	0.39		Peters, Michelle	0.76
Kajs, Larry	0.36		Shermis, Mark	0.63
Gavins, Marva	0.53		Simieou, Felix	0.57
Grigsby, Bettye	0.20		Vesey, Winona	0.90
Morgan, Bryan	0.50		Spuck, Dennis	0.4
Huss-Kessler, Rebecca	0.50		Stockton, Carl	0.7
Pitre, Esrom	0.56		Weaver, Laurie	0.1
Jones, Lisa	0.15		Staples, William	0.4
Kahn, Michele	0.21		Wagner, Paul	0.7
Marquez, Judith	0.32			
Matthew, Kathryn	0.08			

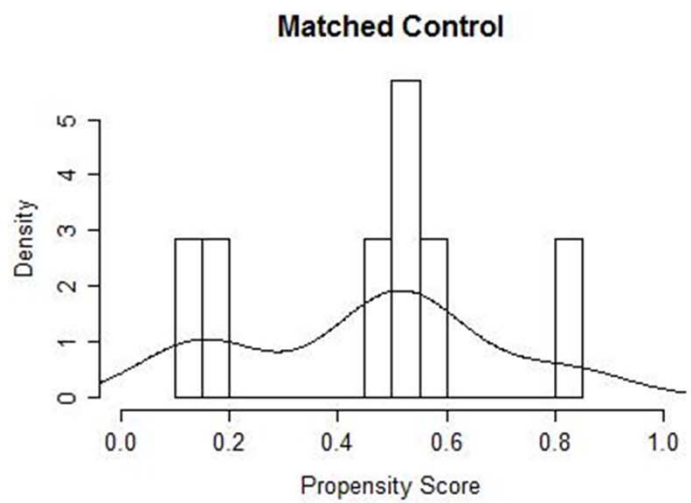
Graphical Representations of Balance



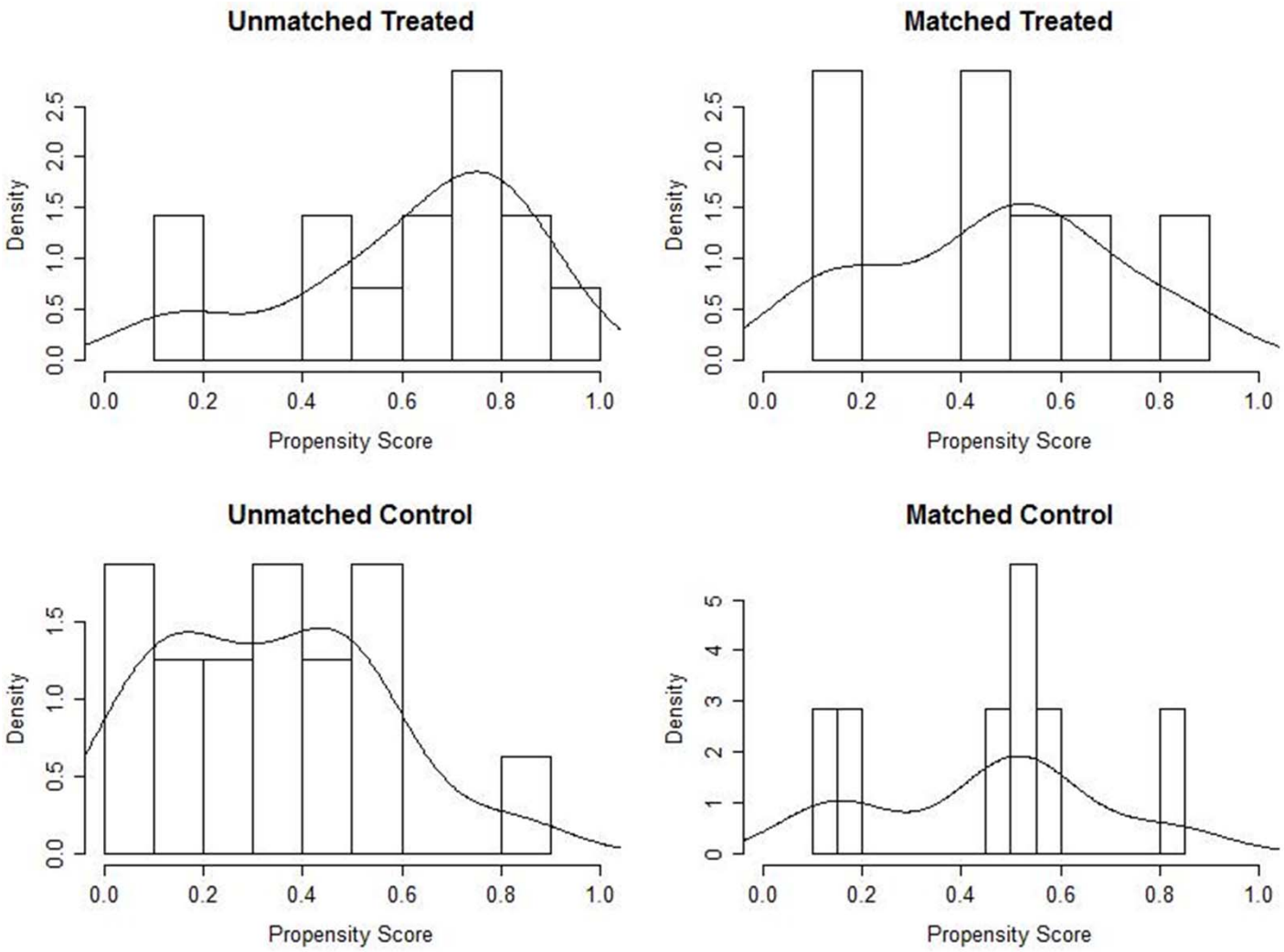


Graphical Representations of Balance





Graphical Representations of Balance



Numeric Representations of Balance

Propensity score matching assumes that, once matched, there are no systematic differences between groups.

	Control		*Intervention*					
	M	*SD*	*M*	*SD*	*t*	*df*	*p*	*d*
Pre-Matching								
π	.33	.22	.62	.24	3.41	28	<.001	1.29
Post Matching								
π	.44	.26	.46	.25	0.14	12	.89	0.08

Cohen's d is a measure of the relationship between two means.

Smaller value = less difference in means = less effect

Larger value = larger difference in means = larger effect

Back To Our (Matched) Data

	Unmatched
Average increase in books read by students not participating in reading intervention:	.06
Average increase in books read by students participating in reading intervention:	.64
Effect size of intervention program (Cohen's *d*)	.85

Cohen's d is a measure of the relationship between two means.

Smaller value = less difference in means = less effect

Larger value = larger difference in means = larger effect

Back To Our (Matched) Data

	Unmatched	Matched
Average increase in books read by students not participating in reading intervention:	.06	.14
Average increase in books read by students participating in reading intervention:	.64	.43
Effect size of intervention program (Cohen's *d*)	.85	.42

Cohen's d is a measure of the relationship between two means.

Smaller value = less difference in means = less effect

Larger value = larger difference in means = larger effect

What We Covered….

- Matching data using propensity scores
- What is a propensity score?
- A need to plan for causal inferences
- National priorities often drive local responsibilities

And, What We Didn't Cover…

- How to create propensity scores
- How to match groups based on propensity sco
- How to evaluate the quality of the match

Thank You!