The Value of Large Scale Data Bases Versus Randomized Experiments for Educational Research

Kenneth A. Frank
Nell K. Duke

Chicago Il, April, 2003.

Direct correspondence to :
Ken Frank
Room 460 Erickson Hall
Michigan State University
East Lansing, MI 48824-1034

E-mail: kenfrank@msu.edu. Phone: 517-355-8538. Fax: 517-353-6393

Full text available at: http://www.msu.edu/~kenfrank/
Abstract

We consider the value of national longitudinal data bases, such as NELS, given the current press for randomized experiments in educational research. The value of randomized experiments is explored, as are challenges to implementation, and we relate to a current attempt to implement a randomized experiment to study effects of teacher’s use of multiple genres on literacy. We then argue for the value of quantifying the robustness of inferences from any design, and apply to analyses of Catholic schools and vouchers. Thus we may be better able to assess and compare the internal validity of inferences from quasi-experimental designs and the external validity from experiments. Finally, we identify key features for extending the usefulness of quasi-experiments or randomized experiments.
What Have We Learned from National Data Bases?

Project Talent, High School and Beyond and the NELS

Factors that can affect changes in achievement, teachers attitudes, and school structures.

National Longitudinal Study of Youth

How high school experiences translate into further opportunities and constraints.

Or bibliography at http://www.nlsbibliography.org/
http://www.pop.psu.edu/data-archive/daman/nlsy1.htm

Early Childhood Longitudinal Study

Factors critical for school readiness.
(West, et al. 2003)
Current Press for Experiments

“The recent enactment of *No Child Left Behind*, and its central principle that federal funds should support educational activities backed by “scientifically-based research,” offers an opportunity to bring rapid, evidence-driven progress — for the first time — to U.S. elementary and secondary education. **Education is a field in which a vast number of interventions, such as ability grouping and grade retention, have gone in or out of fashion over time with little regard to rigorous evidence.** As a result, over the past 30 years the United States has made almost no progress in raising the achievement of elementary and secondary school students, according to the National Assessment of Educational Progress, despite a 90 percent increase in real public spending per student. Our extraordinary inability to raise educational achievement stands in stark contrast to our remarkable progress in improving human health over the same time period — progress which, as discussed below, is largely the result of evidence-based government policies.”

“**Randomized controlled trials** may offer a key to reversing decades of stagnation in American education, and sparking rapid, evidence-driven progress.”

http://www.excelgov.org/displayContent.asp?Keyword=prppcEvidence
Advantages of Randomized Experiment

Statistical

Define:

\( y_i^T \), the value on the dependent variable that would be observed if unit \( i \) were exposed to the treatment;

and \( y_i^C \), the value on the dependent variable that would be observed on unit \( i \) if it were exposed to the control.

Counterfactual:

“it is impossible to observe the value of \( y_i^T \) and \( y_i^C \) on the same unit, and therefore it is impossible to observe the effect of \( t \) [the treatment] on \( i \).”

Random assignment insures

\( E(y_i^T) = E(y_i^T | A=t) \) and \( E(y_i^C) = E(y_i^C | A=c) \).

That is, that the expected value is independent of assignment.

Then \( E(y_i^T) - E(y_i^C) = E(y_i^T | A=t) - E(y_i^C | A=c) \), which can be directly estimated from the data.

Scientific: Reduces likelihood of a confounding variable because those in treatment should be the same on all characteristics as those in the control

Political: Findings considered more robust, policies more stable.
### Differences between Medicine and Education, with Implications for Experiments

<table>
<thead>
<tr>
<th></th>
<th>Randomized Paradigm from Medicine</th>
<th>Education</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Delivery</strong></td>
<td>simple (a pill) (the typical focus of the experiment)</td>
<td>complex (teaching diverse learners)</td>
</tr>
<tr>
<td></td>
<td><em>delivery can be double blind, consistent across physicians</em></td>
<td><em>delivery cannot be double blind, variable by teacher</em></td>
</tr>
<tr>
<td><strong>provider-receiver</strong></td>
<td>physician-patient</td>
<td>teacher to multiple students</td>
</tr>
<tr>
<td>at time of delivery</td>
<td><em>Unit of analysis is patient</em></td>
<td><em>Unit of analysis is teacher or school</em></td>
</tr>
<tr>
<td><strong>Process</strong></td>
<td>Internal to body, perceived to be hidden</td>
<td>External, in interactions, perceived to be easily observed.</td>
</tr>
<tr>
<td></td>
<td><em>accept mystery of science</em></td>
<td><em>Less accepting of mystery of science</em></td>
</tr>
<tr>
<td><strong>lay person information</strong></td>
<td>little knowledge of medicine</td>
<td>everyone is an expert,</td>
</tr>
<tr>
<td></td>
<td><em>authority to physicians</em></td>
<td><em>challenge professional status of teacher, researcher</em></td>
</tr>
</tbody>
</table>
An Attempt to Implement a Randomized Study

Early Literacy Project

| Delivery | complex (teaching diverse learners)  
|          | Highly variable by teacher, with considerable overlap between treatment and control groups |
| provider-receiver at time of delivery | teacher to multiple students  
|                                  | Unit of analysis is teacher or school |
| Process | External, in interactions, perceived to be easily observed.  
|                                   | Monitoring of fidelity (advantage) |
| lay person information | everyone is an expert,  
|                        | some districts reluctant |
Back to the Quasi-Experiment, and Concerns about Confounding

Attendance at Catholic School ($X$)

Math Achievement ($Y$)

Educational Engagement ($V$)

$\lambda_{x,v}$

$\lambda_{x,v} \times \lambda_{y,v}$

$\lambda_{y,v}$
Concerns about Confounding Variables: Internal Validity

Define the impact, $k$, of a confounding variable $v$, on the inference for $x$ as

$$ k = h_{v@y} \times h_{v@x} $$

where $h_{v@y}$ is the hypothetical correlation between $v$ and the outcome, $y$; and $h_{v@x}$ is the hypothetical correlation between the $v$ and the predictor of interest.

Question: How large, exactly, must be the impact of a confounding variable, $k$, to alter a statistical inference?

Assumption: the impact of the confounding variable is maximized (giving maximal credence to the skeptic) – see Frank 2000.

Answer:

$$ k(r_{x@y}) > \frac{r_{x@y} - r^#_{x@y}}{1 - r^#_{x@y}}. \quad (1) $$

Where $r^#_{x@y}$ is a correlation that is just statistically significant:

$$ r^#_{x@y} = \frac{t'}{\sqrt{(n-q-1)+t'^2}}. \quad (2) $$

Thus if $k$ is larger than the TICV($r_{x@y}$) then the impact of the confounding variable alters the original inference. Thus the right side of (1) defines the threshold for the impact of a confounding variable, TICV($r_{x@y}$).

Critic: “But you have not controlled for $v$.”

Researcher: “True, but the impact of $v$ must be larger than the TICV to alter my inference.”
Define $\lambda_{x@}$ = the correlation between $x$ and $y$ in some unobserved sample.

**Question:** If some originally observed cases were replaced, how small must be $\lambda_{x@}$ to alter the inference?

**Assumptions:**
1) Unobserved means and variances equal observed means and variances
   
   $\mu_x^G = \mu_x^G$, $\Delta_x^2 = \sigma_x^2$ and $\Delta_y^2 = \sigma_y^2$,
   
   2) 50% of sample replaced with cases with $\lambda_{x@}$

**Answer:** Inference changes if

$$\lambda_{x@} < 2 \lambda_{x@} - \lambda_{x@}. $$

Called threshold for neutralization by unobserved cases TNUC ($\lambda_{x@}$).

**Critic:** But your sample does not include enough of group $j$!

**Researcher:** True, but $\lambda_{x@}$ for group $j$ would have to be less than the TNUC to alter my overall inference (if half of my sample were replaced with cases from group $j$).
Combining Indices for Internal and External Validity

Alternative expression of $\lambda_{xy}$ relative to $r_{xy}$:

$$\mathcal{D} = \frac{(r_{xy} \lambda_{xy})}{2r_{xy}}$$

$\lambda_{xy} = \text{TNUC}(r_{xy})$

Defined as counterbalancing threshold for unobserved cases, or CTUC($r_{xy}$).

Like the TICV, if the CTUC is large, the inference is more robust, and both

$$0 < \text{TICV} < 1$$
$$0 < \text{CTUC} < 1$$

(assuming $r_{xy}$ is statistically significant and positive)

Combined Threshold of Robustness $\text{CTR}(r_{xy}) = \text{TICV}(r_{xy}) \times \text{CTUC}(r_{xy})$

can then be used to compare results across designs

(e.g., experimental versus quasi-experimental)
### Table 1
Effects, Inferences and Indices of Robustness for Catholic Schools and Vouchers

<table>
<thead>
<tr>
<th>Source</th>
<th>Standard Statistics</th>
<th>Robustness of Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Type of Validity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unobserved Process</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unobserved Quantity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Index</td>
</tr>
<tr>
<td>Morgan (replication of Coleman et al.)</td>
<td>.99 (.33)</td>
<td>3.00</td>
</tr>
<tr>
<td>Mayer et al.$^b$</td>
<td>5.50 (1.38)</td>
<td>3.99</td>
</tr>
</tbody>
</table>

$a$ Note that $r_{\text{achievement covariates}}^2 = .034$ and $r_{\text{catholic school covariates}}^2 = .222$ as estimated by separate analyses of the data (using multilevel modeling software to account for the nesting of students within schools). The corresponding TICV is .031 using equation (8) and $k$ is maximized when $h_{v\text{catholic school}} = .17$ and $h_{v\text{achievement}} = .19$ as in equation (9).

$b$ Mayer results apply only to African Americans; overall effect was not statistically significant.
### Table 2
Comparison of Designs for Assessing Educational Effects

<table>
<thead>
<tr>
<th>Source</th>
<th>Treatment</th>
<th>Unit of analysis</th>
<th>Grade levels$^a$</th>
<th>Design</th>
<th>Sampling</th>
<th>Index of Overall Robustness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morgan (replication of Coleman et al.)</td>
<td>Catholic Schools</td>
<td>School</td>
<td>12</td>
<td>Quasi-Experimental (control for pre-test)</td>
<td>Stratified Random</td>
<td>CTR=$TICV \times CTUC$ $\mathcal{W}, \kappa$</td>
</tr>
<tr>
<td>Mayer et al.$^b$</td>
<td>Vouchers</td>
<td>Student</td>
<td>4-7</td>
<td>Random Assignment</td>
<td>Volunteer</td>
<td>.034</td>
</tr>
</tbody>
</table>

$^a$ when outcome was measured.

$^b$ Mayer results apply only to African Americans; overall effect was not statistically significant.
Implication: Inferences from national data bases compared with experiments

1. Quantification of robustness allows for comparison
2. Measured covariates (especially pre-test) in national data base absorb the impact of other covariates
3. Can address large variety of questions
4. Levels include student, teacher, school, parent

The Shadow of NELS
Extension of National Data Base

1. Full Networks
2. Process, more measurements in finite time
3. Link to local community and district contexts

Implementation of Experiments

1. Control, monitor implementation
2. Apply at school level
3. Inform public, educators, parents, and students of value