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

Kenneth A. Frank
Nell K. Duke


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
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.
1. INTRODUCTION: THE NATIONAL DATA BASE VERSUS THE EXPERIMENT

For decades, educational research has benefitted tremendously from the large scale data bases whose design, collection, and analyses have been supported by the United States government. From Project Talent, High School and Beyond and the NELS data bases we have learned of factors that can affect changes in achievement, teachers attitudes, and school structures Coleman, Hoeffer and Kilgore 1982 Lee and Bryk 1989 Lee and Smith 1995 Lee, Dedrick and Smith 1991. From the National Longitudinal Study we have learned how high school experiences translate into further opportunities and constraints. For example, Griliches 1976 Griliches 1977 shows that the estimate of the return to schooling is lower once controlling for ability, as measure in the NLSY. From the Early Childhood Longitudinal Study we have learned the key factors critical for school readiness West, Denton and Germino-Hausken 2003. Each of these findings leads to a deeper understanding of social processes and may ultimately inform public policy.

Relative to randomized experiments, analyses of national longitudinal data sets have considerable capacity to inform causal inferences and policies. In particular, they can (1) identify those factors that are at least related to particular later outcomes; (2) may reveal factors that had previously been overlooked or underestimated in terms of the relationship to later outcomes, and reveal factors whose relationship has perhaps been overconsidered; (3) may suggest factors to manipulate in experimental designs; (4) may provide information in situations in which manipulation, at least directly, will never be possible (e.g., in looking at effects of lead paint exposure); (5) allow for nationally representative samples, rarely possible in experimental and many other type of research; (6) provide strong 'bang for the buck' -- providing a great deal of data on factors of interest to a great number of different researchers (and policy makers and practitioners) at once (in contrast to most experiments).

And yet recently, the federal government has announced a new emphasis on the randomized experiment. Spurred by the apparent accumulation of knowledge through experimentation in medicine, the federal government has prioritized a similar research paradigm for education (U.S. Department of Education 2002). They claim “Randomized controlled trials may offer a key to reversing decades of stagnation in American education, and sparking rapid, evidence-driven progress.” At the very least the focus on experiments may distract from the
value of the large scale data base. At most it may direct resources away from investment in large scale data bases.

While we agree that experiments have a unique and valuable role in the educational research enterprise, we are concerned with a potential overemphasis on randomized experiments at the expense of continued support for large scale data bases that can inform educational policies. This certainly holds in education because of the difficulty of implementing randomized experiments in school settings because of the complex delivery of education, its multilevel nature and public perceptions of educators. Moreover, to make the case for preserving funding for large scale data bases we will quantitatively compare the bases of causal inferences from randomized experiments versus those from large scale data bases. In particular, we will employ recent advances in sensitivity analysis to consider the robustness of inferences across different designs.

In the next Section we review the strengths of the randomized experiment in terms of the epistemology of the counterfactual. In Section 3 we then consider key limitations to conducting randomized experiments, especially the multilevel nature of schooling. In Section 4, we explore these limitations through a specific example of an attempt to implement a randomized experiment to evaluate literacy practices. This leads us in Section 5 to present indices for the robustness of inferences from the large scale data base relative to those from an experiment. In Section 6 we apply the indices to studies of Catholic schools from national data bases. Finally, in Section 7 we consider the capacity of randomized experiments and quasi-experiments from nationally representative, longitudinal data bases to sustain inferences and inform policy.

2. STRENGTHS OF RANDOMIZED EXPERIMENTS

Assume a policy-maker is considering whether or not to implement or expand an existing educational practice. Furthermore, assume that analyses of extant, non-experimental data, indicate that those who were exposed to the practice had statistically significant greater achievement than those who did not. For example, consider the effect of Catholic schools on educational achievement. If the corresponding effects is large enough to warrant resource expenditures then the policy-maker may make an inference regarding the practice as one cause
of achievement. Correspondingly, the policy-maker may attempt to implement the practice in new contexts, believing it will improve achievement.

Of the many, complex critiques of the above logic (see Cook and Campbell 1979 Cook 2002), a key critique derives from the use of non-experimental data as a basis for a causal inference. Correlation does not equal causation (Holland 1986), we are reminded again and again. In particular, there may be some alternative explanation, some confounding variable, that renders the observed correlation spurious. For this type of example the concern can also be expressed in terms of selection: those who were exposed to the treatment were self-selected on some educational oriented characteristic, and it is this characteristic, not the practice, which produced the observed differences in achievement. For example, those who attended Catholic schools may be more educationally engaged than those who attended public schools.

One can respond by attempting to statistically control (using multiple regression, structural equation models, selection models, etc.) for differences in characteristics of those who were exposed to the practice versus those who were not. Conditional on the controls, those who were exposed to the practice are like others, and therefore differences in achievement can once again be attributed to the practice.

Of course, no matter how exhaustive the list of covariates, there is always the concern that those who were exposed to the practice were different from those who were not on some fundamental characteristic that is related to achievement. That is, that there is still some unmeasured, confounding variable. Even if the researcher had controlled for all of the relevant characteristics there is no way of knowing this to be true, and therefore the causal inference is always suspect.

On the other hand, if the research had been based on a randomized experiment all concerned would have more confidence in the causal inference (Cook 2002). Given large enough samples, randomization insures that those exposed to the practice are similar in most relevant respects to those who were not. Therefore any difference in achievement can be attributed to exposure to the practice (see Rubin 1974; Holland 1986). That is, since a variable must be correlated with both a predictor and outcome to be confounding, and randomization insures no variable is correlated with the predictor indicating exposure to the educational practice, randomization insures there are no confounding variables. Thus one can be more
certain of causal inferences from randomized studies, and more confident in corresponding policies.

Ultimately, one of the greatest values of randomized experiments is in refuting the counterclaims of skeptics. Thus educational policies can be sustained, not subject to “fashion” (U.S. Department of Education 2002). Educational resources can then be efficiently invested in educational practices, culminating in more effective schools. Thus given the perceived value of randomized experiments, educational researchers are called upon to model themselves after those on medicine who epitomize the randomization to effective practice paradigm. Just as medicine is perceived as languishing until it embraced the experimental paradigm, which was finally able to weed out effective practices from ineffective practices, so education should be able to do the same by adopting a similar paradigm.

3. LIMITATIONS OF EXPERIMENTS

Though there is great advantage to the sharing of methods and epistemology across disciplines (e.g., the contributions of meta-analysis to medicine, education, and psychology), there are considerable differences between education and medicine concerning the complexity of the treatment, the level of implementation (corresponding to the unit of analysis), and the level of prior knowledge of the subject. These differences preclude easy transmission of the randomized paradigm from medicine to education.

First, the experimental paradigm is more naturally adapted to the simple implementation of treatment, such as medicine, versus the complex delivery of education. That is, the delivery of a pill or even a simple procedure varies less from doctor to doctor than does the delivery of a curriculum from teacher to teacher. In fact, patient compliance is almost a dirty word in medicine, and medical researchers often resort to non-experimental techniques when compliance varies by type of treatment. Thus the experimental paradigm favors those treatments which are simplest to implement in a time when we appreciate the increasing complexity of the social and cognitive processes of learning.

---

1We refer to the “randomized paradigm from medicine” acknowledging that medical research employs a variety of methods. But the method suggested for education is that of the randomized experiment for a easily delivered treatment.
Second, medical treatments are essentially implemented in a one-to-one relationship between physician and patient. Thus a single physician can implement both treatment and control, and randomization is at the level of the patient. In contrast, a teacher engages multiple students at once and it is unreasonable in most cases to ask the teacher to implement one treatment with one set of students and an alternate treatment with a second set of students. Even if the teacher could implement multiple treatments simultaneously, the students would interact (as encouraged by the social view of education), inducing interference between subjects (violating key assumptions of inference, e.g., Rubin’s SUTVA, 1986). Thus, for most purposes, the minimal level of implementation is the teacher, not the student. This changes the typical scale, and thus expense of any experiment. This problem is further compounded by the increasing appreciation for schools as social institutions. If the best schools are those in which teachers interact and cooperate, then again there is interference among teachers as units within the same school. Thus the most appropriate level of analyses may be the school, further increasing the expense of a randomized experiment. (Of course, some patients are treated in hospitals, but hospital practice is hardly the model for the application of accumulated science.)

Third, the lay person perceives herself to be much more informed about education than medicine. Though patient activist groups and on-line sharing are newly emergent, the level of parental involvement and knowledge about schools is far greater than in medicine. A mother can complain to a teacher that she learned how to read a certain way from her mother, but rare is the mother who could request a pediatric heart surgeon adopt a different technique based on her own experience with such surgery. The point is that we perceive many of the processes of health to be hidden inside the body while we perceive the processes of learning to be more apparent through day-to-day interaction. The implication for experiments is that educational researchers have to work much harder to insure randomization is not undermined by subjects who learn on their own and perceive advantages of certain treatments.

4. AN ATTEMPT TO IMPLEMENT A RANDOMIZED STUDY

*To appreciate these limitations, consider Nell’s experience*
Power Analysis:

As indicated earlier, this study would involve 20 first grade experimental classes, 20 first grade control classes, 20 first grade control-turned-experimental classes, and 20 each of the same groups for second grade (with no overlap between the children in the first grade classes and the children in the second grade classes). Hierarchical power analysis for an effect size .3, intraclass correlation of .1, and \( p \) at .05 indicates the need for 22 classes per condition and 20 students per class to achieve a power of 0.8. However, this power analysis does not take into account the multiple timepoints/assessment points to be used in this study (three per year), the use of multiple outcome measures (9 are planned), or other characteristics of the assessments (as yet unknown). Taking these factors into account we can assume that somewhat fewer classrooms are needed. Twenty classrooms per group, even accounting for some attrition, is justifiable.

Until recently, most quantitative analyses including more than one person affiliated with a given school simply did not account for dependencies among the observations (e.g., Bowles and Gintis 1976; Coleman, et al. 1966 Dreeben and Gamoran 1986; Epstein and Karweit 1983 Kilgore and Pendleton 1993). The analyses employed models that did not represent the nested nature of the phenomena, with negative consequences for statistical inference (Bryk & Raudenbush 1992). Others have met the statistical assumption of independent error terms by analyzing data aggregated to the school level, thereby eliminating the difficulty introduced by dependencies among observations within a single school (e.g., Hannaway and Talbert 1993), by analyzing data from a few schools and accounting for school effects through fixed effects estimation (e.g., Hallinan 1992) by analyzing observations from different schools (e.g., Chew 1992) or by focusing on treatments applied at the level of the individual (*get from Nell*). But although we can learn of commonalities across schools from such data, theoretical and mathematical models built from these data necessarily ignore the complex processes within schools as organizations. In these cases the statistical tail is wagging the theoretical dog.

---

2 We are using recently-developed software for power analysis that is an improvement over traditional power analysis software in taking into account the nested nature of the data. However, this software cannot consider all of these factors simultaneously, nor, to our knowledge, can any software currently available.
5. INDEXING THE ROBUSTNESS OF INFERENCES

5.1 Concerns About Confounding Variables: Internal Validity

Concerns about randomized experiments do not alleviate the fundamental limitation of the non-experimental study – the potential for spurious results attributable to confounding variables. For example, critics of the Catholic school effect have argued that the relationship between attendance in Catholic schools and achievement could be attributed to the higher level of educational engagement of those who attended Catholic schools (Alexander and Pallas 1985). In the absence of random assignment, optimally researchers can measure potential confounding variables and control for them, thus reducing concerns about selection, conditional on the covariates (Rubin 1974). But no matter how many confounding variables are measured and added to a model, and even with random assignment (Cook 2002), there is always the concern that there is an unmeasured confounding variable such that, if controlled for, would alter an inference (where “altered” means going from statistically significant to not statistically significant)\(^3\).

Instead of employing extensive statistical controls and procedures to address concerns about selection, Frank 2000 developed an index of the impact of a confounding variable necessary to alter an inference. This index helps the researcher assess the size of the concern regarding selection, and therefore its effect on inference.

Frank begins by defining the impact, \(k\), of a confounding variable on a statistical inference as the product of two hypothetical correlations: \(k = h_v \cdot h_x\) where \(h_v\) is the hypothetical correlation between the confounding variable, \(v\), and the outcome, \(y\), and \(h_x\) is the

---

\(^3\) We recognize that statistical significance is rarely sufficient for programmatic change. One should of course consider effect sizes, the causal mechanisms and the nuances of implementation in considering programmatic change. But it would also be unusual for an empirical relationship that was not statistically significant to be relied upon as a basis of programmatic change. Therefore statistical significance is essentially a necessary condition, and thus the robustness of the significance is critical to informing the change process.
hypothetical correlation between \( v \) and the predictor of interest, \( x \) (see also Pan 2003). Note that it is assumed that an initial inference has been made and no new data will be analyzed. Thus script is used for terms such as \( \hat{k} \), \( \hat{r}_v \) and \( \hat{r}_x \) that refer to sample statistics that depend on unobserved cases.

Figure 1 uses the concept of impact to express the challenge to the Catholic school effect. Begin with the standard representation of the relationship between \( x \) and \( y \), referred to as \( r_{x,y} \) associated with the arrow at the top of the figure. For ease of presentation we present Frank’s approach in terms of \( r_{x,y} \) (the estimate of correlation between \( x \) and \( y \)) noting that the statistical tests for a regression coefficient and \( r_{x,y} \) are equivalent (Cohen and Cohen 1983).

Next, introduce the concern for the confounding variable in terms of the relationships associated with the confounding variable, \( h_v \) and \( h_y \). The impact is then expressed in terms of the arrows emerging from \( h_v \) and \( h_y \) and converging to represent the product \( h_v \times h_y \) which then impacts on \( r_{x,y} \). For example, the challenge to the Catholic school effect is that the impact of educational engagement would reduce the relationship between Catholic schools and achievement.

The question then is: How large must be \( \hat{k} \) to alter the inference? Frank’s index answers this question directly. First, define \( r# \) as the value of \( r \) that is just statistically significant. That is,

\[
\hat{r#} = \frac{t'}{\sqrt{(n-q-1)+t'^2}},
\]

where \( t' \) is the critical value of a t-distribution on \( n-q-1 \) degrees of freedom. Frank showed that, given the constraint \( \hat{k}=h_v \times h_y \) the impact of a confounding variable is maximized (giving maximal credence to concerns about the impact of \( v \)) when \( h^2_v = h^2_y = h_y \times h_x \). Thus, assuming

\( h^2_v \times h^2_y \times h_y \times h_x = \hat{k} \), Frank showed that if
then the inference regarding $r_{x@y}$ will be altered. Thus the right hand side of (2) defines the threshold for the impact of a confounding variable, referred to as the TICV($r_{x@y}$); if $k^*$ is larger than the TICV($r_{x@y}$) then the impact of the confounding variable alters the original inference. Thus the TICV indicates just how large the threat to an inference must be to alter the statistical, and thus causal inference.

Frank (2000) also extends the simple expression in (2) to a partial correlation, applying to a model containing covariates, $g$. When covariates $g$ are in the model, maximizing the impact yields (see Frank, 2000, page 167, equation 18)

$$TICV(r_{xyg}) = \sqrt{(1-r^2_{xg})(1-r^2_{yg})} \left( \frac{r_{xy}-r^*_{xy}}{1-r^*_{xy}} \right).$$

where $r_{x@}$ is the multiple correlation between $x$ and $g$, $r_{y@}$ is the multiple correlation between $y$ and $g$, and $r_{x@y}$ is the partial correlation of $x$ and $y$, given $g$. Under these conditions, $k^*$ is maximized when

$$k^* = TICV(r_{x@y}) = \sqrt{(1-r^2_{xg})(1-r^2_{yg})} \left( \frac{1-r^2_{x@y}}{1-r^2_{y@g}} \right).$$
Expressing concerns about confounding variables in terms of the TICV shifts the language from the sufficiency of confounding variables to sustain internal validity (Dawid 2000), or the “richness” of existing confounding variables to account for variation in $y$. Stone 1993, to the language of how large, exactly, must be the impact of a confounding variable to alter a statistical inference. It allows the researcher to respond to the critique “But you have not controlled for $v$.” with “True, but how large must be the impact of $v$ to alter my statistical inference?” Thus the TICV informs the discussion of statistical and causal inference and, ultimately, the accumulation of knowledge from quasi-experiments. But the TICV addresses the robustness of inference only with respect to concerns about internal validity and confounding variables. Therefore, in the next sub-section we apply a similar hypothetical framework to develop an index of the robustness of an inference to concerns about sample representativeness and external validity.

5.2 Concerns about Unrepresented Populations: External Validity

Of course, to accommodate the lack of internal validity of quasi-experimental designs, many have recently advocated the value of randomized experiments. But in employing experiments there is the potential loss of external validity by sampling in a specific context in which a particular treatment is being considered. The first response to such a concern should be to estimate and report interaction effects. But even if one tests for interactions with some or all measured covariates, there is always the concern that the treatment could interact with some untested combination of covariates or some unmeasured covariate. Implied is that the overall inference could be altered if there were an un-designated sub-population (for which the treatment effect was weaker) that was more represented in the sample. Thus external validity and the causal inference are compromised, suggesting caution in basing action in new contexts on the causal inference.

Of course, representativeness is always violated to some extent (Little 2000; Cook, 2002). In the logic of robustness indices, the question then is, “If the original sample were made more representative by substituting cases from an unobserved sample, how small must be $\lambda$.@
the correlation between $x$ and $y$ in an unobserved sample, to alter the inference from $r_{xy}$?"

To focus on differences in the relationship between $x$ and $y$ between the observed and unobserved sample, begin by assuming that $\delta_x^2 = s_x^2$ and $\delta_y^2 = s_y^2$, that is that variances of $x$ and $y$ for the unobserved sample equal those of the observed sample. This is a typical assumption of statistical inference that would be applied to the combined data. Furthermore, assume that $G_x = \bar{\delta_x}$ and $G_y = \bar{\delta_y}$, that the means in the unobserved sample are equal to those in the observed sample.

While not an assumption of regression, concerns about the external validity do not pertain to differences in mean values of $x$ and $y$ (and any differences in means could be accounted for by simply including in the model an indicator of whether the case were originally observed or not).

Next, define $n$ as the sample size for the unobserved cases and $n'$ as the number of cases in the original sample that are preserved when some of the originally observed cases are replaced with unobserved cases. Also assume that $n' + n = n$, thus preserving the overall sample size and degrees of freedom for inference. Given $\delta_x^2 = s_x^2$, $\delta_y^2 = s_y^2$, $G_x = \bar{x}$ and $G_y = \bar{y}$, the correlation in the combined data, is simply the weighted average of the observed and unobserved correlations: $R_{x\@} = (n' r_{x\@} + n r_{x\@})/(n)$. (Note that the capital script letter refers to a quantity defined by both observed and unobserved cases.)

If we require that $n = n' = n/2$ then $r_{x\@}$ and $\lambda_{x\@}$ contribute equally to $R_{x\@}$ and the impact of the originally unobserved cases on $R_{x\@}$ can be directly attributed to $\lambda_{x\@}$ and not to disproportionate substitution (e.g., the inference regarding $x_1$ is unlikely to be altered if only one case is replaced). Assuming $n' = n = n/2$ yields $R_{x\@} = (r_{x\@} + \lambda_{x\@})/2$. To obtain the value of $\lambda_{x\@}$ that reduces $R_{x\@}$ to the threshold of statistical significance, set $R_{x\@} = r_{x\@}$ and solve for $\lambda_{x\@}$.
\[ h_{x@} = 2 r_{x@}^* - r_{x@} \]

Thus if \( h_{x@} \) is less than \( 2 r_{x@}^* - r_{x@} \) then the original inference based on \( r_{x@} \) will be altered when half of the observed cases are replaced with originally unobserved cases with sample correlation \( h_{x@} \). By using (1) to substitute for \( r_{x@}^* \) we can obtain the magnitude of the correlation in the unobserved cases required to make the combined data just statistically significant. Thus \( 2 r_{x@}^* - r_{x@} \) defines the threshold for neutralization by unobserved cases, referred to as the TNUC (\( r_{x@} \)).

Using partial correlation or regression coefficients, the TNUC(\( r_{x@} \)) is easily generalized to models with covariates. Begin by assuming that \( h_{x@} = r_{x@} \) and \( h_{y@} = r_{y@} \) that is, that \( x \) and \( y \) are correlated with \( g \) in the unobserved sample just as they are in the observed sample. Were this not the case, then we return to the concerns about internal validity. Then \( h_{x@|g} = 2 r_{x@|g}^* - r_{x@|g} \)

\[ h_{y@|g} = 2 r_{y@|g}^* - r_{y@|g} \]

The expression is essentially unchanged because the sampling distribution and test statistic for a partial correlation are identical to those of a zero order correlation except for a correction for degrees of freedom, which is represented in \( r_{x@}( \text{Weisberg 1985}) \).

Like the TICV, expressing concerns about external validity in terms of the TNUC shifts the language from the possibility of non-additive effects to specification of how small, exactly, must be \( h_{x@} \) such that if half of the observed cases are replaced by unobserved cases the original inference would be altered. Thus the TNUC allows the research to respond to the critique “But your sample does not include enough of group \( j \)” with “True, but \( h_{x@} \) 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 \( j \).” Because the TICV and TNUC correspond to concerns about internal and external validity they provide complementary information about the robustness of an inference. In the next sub-section, we combine the two into a single index of robustness which can be used for comparing results across different types of designs.
5.3 Combining Indices for Internal and External Validity

The TICV and TNUC can be combined to index concerns about internal validity (potentially confounding variables) and external validity (unobserved cases). Begin by combining representations for unmeasured confounding variables and unobserved cases by including $k$ in the equation for $\hat{k}$. That is, replace $r_{x@y}$ in (2) with $\hat{R}_{x@y} = (r_{x@y} + k)/2$:

$$\hat{R}_{x@y} = \frac{r_{x@y} + k}{2}$$

(5)

Setting equal to $r_{x@y}$ then represents how statistical inferences are impacted by confounding variables (as represented by $k$) and correlations in unobserved samples (as represented by $r_{x@y}$).

To compare the two forms of robustness, convert $r_{x@y}$ to the same range as $k$ and impacting inference in the same direction as $k$. In particular, define $D = (r_{x@y} - k)/2r_{x@y}$. Then, for $r_{x@y} > 0$ the greater $D$, the more likely is the inference to be altered. Furthermore, to focus on external validity let $k = 0$ or, equivalently, $r_{x@y} = \text{TNUC}(r_{x@y})$. Then

$$D = \frac{\hat{r}_{x@y} - k}{r_{x@y}}$$

(6)

Expressed in terms of $r_{x@y}$ the right hand side of (6) defines the counterbalancing threshold for unobserved cases, or CTUC$(r_{x@y})$. Frank (2003) shows that given the focus on the impact of the unobserved cases necessary to reduce $r_{x@y}$ below the level of statistical significance, $0 < \text{CTUC}$
(r_s) < 1. Thus CTUC (r_s) has the same range as the TICV(r_s) and, like the TICV(r_s), the larger the CTUC (r_s) the more robust the inference.

Using \( D \), equation (5), containing \( k \) and \( \gamma_s \) can then be reexpressed as

\[
\gamma_{x \cdot y} = \frac{r_{x \cdot y} (1 - D) - k}{1 - k} = r''_{x \cdot y} .
\]  

Solving (7) for either \( k \) or \( D \) in terms of the other yields

\[
D = 1 - \frac{r''_{x \cdot y} (1 - k) + k}{r_{x \cdot y}} ,
\]  

and

\[
k = \frac{r_{x \cdot y} (1 - D) - r''_{x \cdot y}}{1 - r''_{x \cdot y}} .
\]

Drawing on the earlier expansions, these simple expressions directly generalize to models with covariates \( g \).

Using (7) or (9) an expression of overall robustness can be obtained by integrating over \( k \) or \( D \) respectively. Frank (2003) shows that this yields TICV(r_s)×CTUC (r_s) as a combined threshold of robustness (CTR[r_s]). Thus the (CTR[r_s]) is an overall measure of robustness, which can then be used to compare results across designs (e.g., experimental versus quasi-experimental).
The findings of James Coleman and colleagues (Coleman, Hoffer and Kilgore 1982, Hoffer 1985) drew on exemplary large scale, government funded data bases (High school and Beyond and NELS). Nonetheless, their findings sparked methodological and policy controversy that persist (Alexander and Pallas 1983, 1985; Goldberger and Glen G. Cain 1982 Greene, Peterson and and Du 1998; Howell and Peterson 2000; McEwan 2000; Morgan 2001). The central finding of Coleman and colleagues was that students who attended Catholic schools had higher, and statistically significant, levels of mathematics achievement in twelfth grade than students who attended public schools. The corresponding causal inference is that Catholic schools educate students better than public schools, which was used to support the policy that students should be given vouchers to attend any school: public, private, or religious (Coleman, Hoffer and Kilgore).

Initial critics of Coleman’s inferences charged was that students who selected themselves into, and were admitted to, Catholic schools were more engaged in education and already had higher levels of achievement than students who attended public schools (e.g., Alexander and Pallas 1985). This is the typical concern regarding the internal validity of an inference from a non-experimental design. Coleman et al attempted to address the concern of selectivity by drawing on their longitudinal data to use measures of achievement in tenth grade as critical covariates for twelfth grade achievement.

Coleman’s analyses have been replicated and extensively reevaluated (Altonji, Elder and Taber 2000; Figlio and and Stone 1999; Gamoran 1996; Goldhaber 1996; Grogger 2001). Arguably the most current and methodologically comprehensive can be found in Morgan (2001). Morgan reproduces Coleman’s analyses, defining the outcome as a difference score between twelfth and tenth grades and then controlling for tenth grade achievement as a covariate (Morgan also explores models which include propensity scores and possible mediating constructs such as educational expectations and parental involvement, course-taking, climate and track).

Coleman was also critiqued for failing to account for the clustering of students within schools which now is typically accounted for with multilevel or random effects models – see Lee and Bryk 1989 – and which will be employed in the analyses presented here)
Essentially confirming Coleman’s findings, Morgan found that the overall effect of Catholic schools on math achievement was statistically significant, controlling for tenth grade math achievement and family background and demographics (a set of fourteen variables). The regression coefficient for Catholic versus public schools was .99 with t-ratio 3.0 and corresponding $r_{\text{Cath} \@ \text{Math}}$ of .097. These results are reported in Table 1. (Morgan’s standard errors reflect dependencies associated with the nesting of students within schools – see note to Morgan’s Table 1, page 345 – and thus the sample size is based on the number of schools, approximately 973. Using multilevel modeling software of Raudenbush et al, 2002, the regression coefficient was 1.0 with standard error .34 and t-ratio of 2.96, confirming Morgan’s results.)

Concerns about internal validity in Morgan’s analysis should be limited, because he controlled for a pre-test. But we cannot be certain that internal validity is completely satisfied. For example, the covariates did not include parental educational expectations and intended engagement of parents, both of which might be higher for those students who chose to attend Catholic schools. Let $\text{par}$ be parental expectations/engagement, an example of a robustness question is then, how large must be $\lambda_{\text{par} @ \text{Math} | g}$ and $\lambda_{\text{par} @ \text{Cath} | g}$ to alter the inference regarding the effect of Catholic schools? The answer given by the TICV($r_{\text{Cath @ Math} | g}$), is that the product must be .035. Therefore if $\lambda_{\text{par} @ \text{Cath} | g} = \lambda_{\text{par} @ \text{Math} | g}$ then each would have to be about .189 to alter the inference (note that the impact is maximized when $\lambda_{\text{par} @ \text{Cath} | g}$ is slightly less than $\lambda_{\text{par} @ \text{Math} | g}$ – see note to table 1).

The TICV highlights the power of the ANCOVA-like design data bases such as NELS. On one hand, .189 is a small to moderate effect in social sciences (Cohen and Cohen 1983). On the other, the effect must be net of the controls already in the model, in particular tenth grade achievement. Thus the zero order correlations of parental expectations/engagement with attendance at Catholic school and twelfth grade math achievement must certainly be larger than .187. That is, the covariates likely absorb some of the impact of parental expectations/involvement (Frank 2000).
Because the data in NELS:88 are nationally representative (when properly weighted) and the data set is large with considerable variation in student demographic characteristics, concern about external validity is not as great as concerns about internal validity. The overall average effect is likely to be similar to that in a comparable sample, and interaction effects can be directly estimated (for example, Morgan makes the point that the Catholic school effect is greatest for those with the highest propensity to attend Catholic schools). This reflects the strong external validity of the quasi-experimental design using nationally representative data. Nonetheless, the TNUC(r_{Cath \@ Math}) can express the robustness of the inference to any additional concerns about external validity. In this case the TNUC(r_{x \@ | g})=.030 with CTUC (r_{x \@ | g}) of .346. Thus if half the NELS:88 sample were replaced with observations for which $h_{Cath \@ Math | g} \leq 0.030$ (or that $r_{Cath \@ Math | g} - h_{Cath \@ Math | g}$ were about 35% of its range), the inference regarding Catholic schools would be altered. Thus the TNUC(r_{Cath \@ Math | g}) quantifies the robustness of the inference regarding the effect of Catholic schools with respect to the assumption of additive, constant, effects.

Coleman et al’s findings triggered policy debates about vouchers and Catholic schools (e.g., Chubb and Moe 1990). The proponents of vouchers argue that if Catholic schools are more effective than public schools then students should be allowed to apply their educational dollars to any school. At the societal level, the belief is that developing a market for schools will improve schools (Friedman and Friedman 1980 Greene 2000). But policy has been slow to follow this logical inference, partly because of concerns regarding the robustness of the causal inferences from quasi-experimental results (Alexander and Pallas, 1983, 1985; McEwan 2000).

To address the limitations of quasi-experimental designs, Paul Peterson and colleagues have recently assessed the effects of vouchers using randomized experiments (e.g., Greene 2000; Howell and Peterson 2000; Mayer, et al. 2002; Wolf 2000; see McEwan 2000, for a review). In varying ways, these researchers have capitalized on relative over-interest in vouchers to randomly assign potential voucher users to be offered vouchers or not. Thus random assignment increases the likelihood that there are no confounding variables, and therefore any causal inference more robust.
The findings of Mayer, Peterson, Myers, Tuttle and Howell are exemplary of many of the studies (Greene 2000). They analyzed the effect of vouchers offered to students in grades 4–7. The sample was obtained from students in New York City who were eligible for the free lunch program. Approximately ninety percent of the sample were African American or Hispanic.

As in other studies, Mayer et al found that vouchers had their strongest effects for African-American students. In particular, the difference (in the third year of the study) of 5.5 national percentile ranking (NPR) points between those who were offered vouchers and those who were not was statistically significant with a t-ratio of 3.99 (t-ratios were not reported by Mayer et al, but this t-ratio was calculated given the sample size of 660 in the treatment and 327 in the control obtained by multiplying the percentages of each group in each condition, as reported in Table A-1 of Mayer et al, by the total n for each treatment. A pooled standard deviation of 20.45 was determined from the ratio of reported mean differences and effects sizes.) These results are reported in the second to last row of Table 1.

Of course, because this study was conducted on selected populations in primarily urban areas, there is considerable concern about external validity. That is, does the inference apply to other types of students or in other geographical areas? This is the typical concern regarding the external validity of an inference from an experiment. Phrased in terms of robustness, what must be \( h_{voucher@PR} \) such that if an alternate sample were substituted for half of the current sample, the inference would be altered?

Calculation of the TNUC(\( r_{voucher@PR} \)) quantifies the concern about external validity. In fact, the TNUC(\( r_{voucher@PR} \)) is -.002 with CTUC(\( r_{voucher@PR} \)) of .508. Thus if half of the New York African American sample were replaced with a sample in which \( h_{voucher@PR} \leq -.002 \) (about half the maximal range), the inference regarding the effect of vouchers on achievement would be altered. Such an alternative sample could come from students outside of heavily urban areas, from students in different cities, or students of different composition. In fact, \( r_{voucher@PR} \) for Latino students (the second largest group in the study) was -.033, and the overall effect across the whole sample was not statistically significant (even with the combined sample size over 2,000). As was the case for analyses of the NELS:88 data, there may still be debate regarding
the robustness of the inference. But the key issue is the terms of the debate, at least as it applies to statistical inference, have now been quantified.

Given random assignment to treatment conditions in the Mayer et al study, there is less concern about internal validity than in the quasi-experimental design. Nonetheless, there is still concern that there may be some violation to internal validity that could alter the inferences (Mayer et al., employ instrumental variables to address the concern). Calculation of the TICV($r_{\text{voucher \& NPR}}$) indicates that a confounding variable would have to have an impact greater than .068, with component correlations of about .261, to alter the inference. Such a moderate correlation for ($r_{\text{voucher \& NPR}}$) is unlikely given random assignment. Therefore either ($r_{\text{\& NPR}}$) would have to be very large or the primary threat to inference concerns external validity as quantified by the indices of external validity.

Evidence from NELS:88 and the voucher experiments differ in many critical ways. As reviewed table 2, the studies differ in terms of the treatment evaluated, unit of analysis, the students’ grade levels, study design and sample. To establish internal validity, Morgan (and Coleman et al) used statistical adjustment in a quasi-experimental design, while the results from Mayer et al’s experiments are reported as simple differences in means. Regarding external validity, the NELS:88 data, based on a random sample, are strongly generalizable to high school students in the United States in the early 1990's. In contrast, the Mayer et al data are based on volunteers from New York. To the extent that the volunteers are similar to others in New York, inferences apply to New York. To the extent that the volunteers are similar to those in other urban districts, the inferences are generalizable to a super population of urban students.

Qualitatively, we can say that the analyses of the quasi-experimental and experimental designs support similar conclusions – Catholic/alternative schools have small but discernable effects on achievement. But though the results are based on different designs, the robustness can be compared quantitatively using CTR($r_{c\&m}$) and CTCC**($r_{c\&m}$). As shown in table 2, the CTR($r_{\text{Cath \& Math}}$) for the Morgan analyses of the NELS:88 data is .011, and the CTR($r_{\text{voucher \& PR}}$) for Mayer et al experimental data is .034. Thus the inferences from the analyses of the
experimental design are more robust than those from analyses of the quasi-experimental design. Note however that CTR($r_{\text{voucher \@ NPR}}$) was calculated for only the sub-sample of African Americans. Results from the whole sample were not originally statistically significant, making the CTR($r_{\text{voucher \@ NPR}}$) originally undefined.

7. DISCUSSION

Though the experimental paradigm appears to offer stronger inferences than the quasi-experiment, it may be difficult to conduct experiments in educational research relative to medicine. “Treatments” in education are considerably more difficult to implement than those in medicine, and it is more difficult to establish a placebo; learning occurs in a classroom, instead of a one-to-one context, and therefore the classroom (or even the school) is a relevant unit of analysis. The public presumes more knowledge about education and therefore is less likely to participate in a randomized experiment. Thus, as in *Nell’s example*, analyses may ultimately need to treat experimental data as a quasi-experiment.

The problem then returns to the lack of internal validity in quasi-experimental studies such as NELS. But how large a problem is this? Should it really undermine our inferences? In this example explored above, the effect of Catholic schools was statistically significant but challenged because it was not conducted as a randomized experiment. But here we ask whether the implied causal inferences inform the likely consequences of action. In particular the inference regarding Catholic schools from the NELS:88 data would have been altered if there were a confounding variable such that $h_{v_{\text{inh}}} \times h_{v_{\text{inh}}} > .035$ (with component correlations of about .19). Thus the correlations would each have to be small to moderate to alter the inference, and this is controlling for a pre-test. Now we know that the inference is not extremely robust, nor is it flimsy. Once we know this, is an experiment really necessary? Indeed, we learn little more from the experiment than we knew from the quasi-experiment - the effect is modest, the inference moderately robust (although concern about external validity is stronger in the experiment).

This is not to say that we cannot learn from experiments, or that experiments do not have
their political value. Randomized experiments in education have helped establish the effects of Head Start and small classes (Finn and Achilles 1990), at least in the early grades (Hanushek). But if experiments are to contribute greatly to educational research, they will require the resources to be conducted at the class level or school level, educational researchers will have to be sophisticated about considering implementation and administering a placebo, and we will have to insure the complete participation of our subjects.

And yet, just as our color palettes can be subdued by black and white film, our conception of causal processes can be inhibited by research designs such as NELS. First, to achieve national representation, NELS sampled only a subset of students and teachers from each school. Lost then are the full networks, and therefore representations of social systems that affect how schools function as organizations and the culture of students’ lives (cf. Bidwell 2001; Coleman 1961; Frank, Topper and Zhao 2000 McFarland 1999).

Second, NELS sampled students intermittently, thus precluding understanding of causal processes as a sequence of events. Thus the NELS paradigm stands almost in opposition of the type of detailed career trajectories of historical analyses. With the lack of contextual and temporal nuance most analyses of NELS suffer from the overemphasis on the variable, and not the process (Abbott).

Third, because student and high school identifications are masked, analyses of NELS type data cannot be understood relative to immediate community contexts. Principals may report on rough characterizations of district policies, but there is not information regarding the inevitable politics through which those policies emerged. Thus schools are studied as almost stand alone institutions, which they almost certainly are not.

The educational research community continues to address these limitations through creative and alternative designs. Charles Bidwell, Milbrey McLaughlin, Maureen Hallinan, Jacclyn Eccles and others study full social networks of teachers and students. Mike Czikamahilai and others sample frequently to capture sequences of events. Others, including Joyce Epstein, Pamela Quiroz and Kathryn Borman study the school relative to its local context.

But as we seek to push past the NELS paradigm, we should appreciate some intrinsic advantages that will be difficult to recover. First, NELS-like data are nationally representative, reducing concerns about external validity. In contrast, researchers pursuing alternatives as
individuals are unlikely to attain such representation. Second, NELS-like data are longitudinal, thus allowing for critical control of pre-tests, making causal inferences more robust. Third, NELS-like data bases contain data at multiple levels, from student achievement to teacher and parent characteristics to principals. This allows for estimation of multiple causes of effects, as well as interactions, reflecting the complex nature of education. Fourth, NELS-like data bases draw their questions from multiple disciplines, including psychology, sociology and economics, and thus can facilitate interaction between disciplines as well as sustain a multi-disciplinary understanding of a given phenomenon.

There are many limitations to the NELS paradigm. But as we consider addressing these limitations, either with experiments or new designs, we must consider the value, long fought for, of the nationally representative, longitudinal data bases we have.
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>$\hat{s}_1$ (se)</td>
<td>$t (\hat{s}_1)$</td>
</tr>
<tr>
<td>Morgan (replication of Coleman et al.)</td>
<td>0.99 (.33)</td>
<td>3.00</td>
</tr>
<tr>
<td>Mayer et al. (^a)</td>
<td>5.50 (1.38)</td>
<td>3.99</td>
</tr>
</tbody>
</table>

\(^a\) Note that $r^2_{achievement@covariates} = 0.034$ and $r^2_{catholic school@covariates} = 0.222$ 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 0.031 using equation (8) and $k$ is maximized when $r_v@catholic school = 0.17$ and $r_v@achievement = 0.19$ as in equation (9).
Mayer results apply only to African Americans; overall effect was not statistically significant.
<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>0.011</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>0.034</td>
</tr>
</tbody>
</table>

\(^a\) when outcome was measured.

\(^b\) Mayer results apply only to African Americans; overall effect was not statistically significant.
References


81, 945-970.


