The Canadian Journal of Program Evaluation Vol. 22 No. 2 Pages 57–68
ISSN 0834-1516 Copyright © 2007 Canadian Evaluation Society

USING PUBLIC DATABASES TO STUDY RELATIVE PROGRAM IMPACT

Tarek Azzam
Christina Christie
Claremont Graduate University
Los Angeles, California

Abstract: Evaluators are under increasing pressure to answer the "compared to what" question when examining the impact of the programs they study. Program contexts and other restraints often make it impossible to study impact using some of our more rigorous methods such as randomized control trials. Alternative methods for studying impact under extreme contextual constraints should be explored and shared for use by others. This article presents a method for studying program impact using existing public datasets as a means for deriving comparison groups and assessing relative impact.

Résumé : On exige de plus en plus des évaluateurs qu'ils répondent à la question « en comparaison à quoi? » lorsqu'ils analysent l'incidence des programmes qu'ils étudient. Le contexte des programmes et d'autres contraintes rendent souvent impossible l'utilisation de certaines de nos méthodes les plus rigoureuses, comme les essais randomisés, dans l'étude de leurs impacts. Il est impératif d'analyser et de diffuser des méthodes de rechange qui en permettent l'étude en dépit de contraintes contextuelles extrêmes. Cet article propose une méthode qui s'appuie sur des banques de données publiques disponibles pour établir des groupes de référence et déterminer l'incidence relative des programmes.

It has been argued that randomized control trials (RCTs) are the most effective method available for measuring the impact of educational programs (National Research Council [NRC], 2002). Randomized control trials require the random assignment of individuals into a treatment group (the group receiving the intervention or program) or a control group (the group not receiving the intervention or program) (Kirk, 1995). Random assignment reduces most threats to internal validity (e.g., selection bias, maturation, history) (Shadish, Cook, & Campbell, 2002), and the use of a control group provides data to evaluate the relative performance of the treatment group,

Corresponding author: Tarek Azzam, 175 East 12th Street, Claremont, CA 91711 USA; <tazzam@ucla.edu>

helping to answer the “compared to what?” evaluation question when measuring program impact.

Although there are many advantages to implementing RCT designs when assessing program impact, there are also shortcomings. Two major criticisms of RCTs are their cost relative to other evaluation study designs and withholding program services to a group that could potentially benefit from the intervention. Scriven (2004), for example, argues that the addition of a control group requires that the evaluation budget be automatically doubled. In addition to cost, it has been argued that withholding services from control group participants can, in some cases, be unethical (NRC, 2002). In addition, the confines of a given context can also make it difficult to randomize participants. This is often true in K-12 educational settings where it is very difficult to randomly assign students to a particular program. These challenges often make the implementation of RCTs unfeasible for studying educational program impact.

Quasi-experimental designs are seen as potential alternatives when RCT designs are not feasible. Quasi-experiments do not require random assignment and often involve the use of a comparison group,[1] as opposed to a control group. However, even quasi-experiments can be difficult to implement, particularly when there are limited resources (financial, human, and time) available for the evaluation. When resources are tight, it is important for evaluators to find alternative means for addressing questions of relative value, instead of abandoning the activity altogether.

When faced with these challenges, we have attempted to find an alternative method of defining a comparison group that is both financially feasible and time efficient. Our solution was to define comparison groups using existing large-scale public and semi-public databases. Many of these databases offer individual-level data on variables that are relevant when measuring the impact of social and educational programs (Table 1). In addition, most of these databases contain demographic information (or other key individual traits) that can be used to create a matched comparison group.

As seen in Table 1, Canada has several public databases available for use by program evaluators. Statistics Canada <www.statcan.ca> serves as one source offering a plethora of public information. Found at this website is an overview of statistical information on Canada’s people, economy, and governments. For an overview of available data, tables can be searched by subject, province or territory, or metropolitan area, or by entering keywords in a search window. Alternatively, tables are listed alphabetically. Also listed are links to other Canadian public databases.

Our evaluation work is conducted primarily in the United States education sector. For our purposes, we find public databases such as the United States National Center for Educational Statistics (NCES) databases very useful. NCES contains variables related to university expectations from a nationally representative student sample. Thus, to illustrate, an evaluator studying the impact of a smaller local program designed to increase students' university-going expectations could use local program demographic data (e.g., ethnicity, socioeconomic status, gender, and location) and the demographic data from NCES to create a comparison group. Local program students' reported post-secondary expectations can then be compared to the re-

Table 1
Sample Databases Related to Education and Health

Statistics Canada	
Website:	<www.statcan.ca/>
Sample Datasets:	*Education, training, and learning*
	Elementary-Secondary Education Statistics Project
	Postsecondary Student Survey
	International Adult Literacy and Skills Survey
	National Longitudinal Survey of Children and Youth
	Post-Secondary Education Participation Survey
	School Leavers Survey
	Survey of Canadian Attitudes toward Learning
	Youth in Transition Survey
	Children and youth
	National Longitudinal Survey of Children and Youth
	Adult and Youth Recidivism in Canada Special Study
	National Alcohol and Drug Survey
	Corrections Key Indicator Report for Adults and Young Offenders
	Youth Smoking Survey
	Youth Custody and Community Services
Access:	Public with application process
Canadian Education Statistics Council	
Website:	<http://www.cesc.ca/>
Sample Datasets:	Financing education systems
	Early childhood development and learning
	Information and communications technologies (ICT) in schools
	Secondary school graduation
	Adult education and training
Access:	Public with application process

National Center for Educational Statistics (United States)

Website:	<http://nces.ed.gov/> <http://nces.ed.gov/dasol/>
Sample Datasets:	Early Childhood – EC Crime and Safety Surveys - CSS Current Population Survey, October - CPS Data on Vocational Education - DOVE Education Finance Statistics Center - EDFIN Education Longitudinal Study of 2002 - ELS High School and Beyond - HS&B High School Transcript Studies - HST National Household Education Survey - NHES National Longitudinal Study of the H.S. Class of 1972 - NLS-72 Private School Survey - PSS Rural Education School District Demographics - SDDS
Access:	Public

National Longitudinal Survey of Adolescent Health (ADD Health) (United States)

Website:	<http://www.cpc.unc.edu/projects/addhealth>
Sample Datasets:	Self-Efficacy Survey Feelings Scale Relations with Parents Survey Friends Survey Delinquency Scale Survey Neighborhood Survey Expectations, Employment, Income Survey
Access:	Public with application process

ported post-secondary expectations of students in the matched group derived from the NCES. This comparative analysis would provide the evaluator with some relative indicator of how students in the local program view the possibility of going to university relative to other students with similar characteristics.

POINTS TO CONSIDER WHEN USING PUBLIC DATABASES

In any evaluation study, the relative costs and benefits of the study design and methods must be weighed. Most public databases are inexpensive to use and can be accessed almost immediately. Costs related to acquiring data from many public databases are often minimal, if not free (Bertrand, Mock, & Franklin, 1981; Stewart, 1984; Webb, Campbell, Schwartz, & Sechrest, 1966). Additionally, gaining access to data can be instant (especially if online access is allowed) or in some cases may require a few days or weeks.[2] Thus, when money and time are limited, public databases can be a desirable option for evaluators looking to understand relative program impact.

It is important to recognize that there are some considerable limitations to using public databases that must be considered before choosing to use them as part of an evaluation study. Being limited when selecting both comparison groups and program outcomes to only the variables included in the public database is one serious restriction of using public datasets. Imagine for a moment that a particular database contains information on students' ethnicity and grade level, but no additional demographic information such as socioeconomic status or parents' education level. The omitted information may well be necessary to select an optimal comparison group when studying a particular program, thus posing a significant problem for the evaluator. There are many other instances where the specific data included in a particular public database would restrict its use when examining program outcomes. For instance, when comparing students' substance use after being involved in a prevention program, an evaluator would be limited to drug comparisons as measured and defined by the public database.

Related to the limitations of selecting comparison groups and outcomes, using public databases may also restrict the set of potential evaluation questions that can be posed to study a program. It has been well established that it is preferable to determine evaluation questions based on stakeholders' information needs (e.g., Patton, 1997; Weiss, 1998). However, when using public databases, the range of possible evaluation questions are limited to the available comparison data (rather than by stakeholders' needs). Educating stakeholders up front about the data available in a particular public database allows those stakeholders to be active participants in determining whether their information needs will be met if a public database is used to understand relative program impact. Thus, it is critical for an evaluator to know exactly what information a public database contains and to convey that information to program stakeholders in a form that is meaningful to them, prior to developing evaluation questions.

As mentioned, public databases can be quite useful when time and money are tight, yet their use admittedly dictates some of the questions asked and methods used in an evaluation study. It is critical, however, that the constraints related to using public databases be restricted to impacting the evaluation rather than the program. That is, using public databases for evaluation purposes should not undermine or be used to determine the program's initial conceptual framework, target population, stated objectives, or implementation. As such, a program's target population should not be changed to create a better match with an existing dataset's population, for example. We feel that this point, which we recognize as extreme, is important to raise as a cautionary note.

Before committing to using a public database, the evaluator should consider the purpose of the database, the accuracy of the data, the sampling strategy employed to collect the data, and the time (date) when the data were collected. Understanding the original purpose of the public database provides the evaluator with information on past use of the database, as well as some insight into how it may best be used for the current evaluation study. When considering the accuracy of a database, the evaluator must determine and understand how the data were collected and recorded and what, if any, changes occurred while the data were being collected. This will help the evaluator avoid any major misinterpretations of database information, and also provide the evaluator with an understanding of the procedures necessary for data collection in his or her specific study. For example, if the database contained survey data collected from students through home interviews, then the evaluator should also attempt to collect survey data using home interviews to minimize location-related biases. How often data were collected is also of concern. It is ideal to collect data as frequently as was done in the database study. For instance, some databases contain data that were collected annually, while others contain data collected once every three to five years. In addition to data collection procedures, sampling is also an important factor to consider when selecting a database. The evaluator must know the population sampled in the database study, the criteria for selection, and—if possible—attrition rates. Sampling information about the candidate database will help determine whether it is appropriate for use in a particular evaluation study. So that proper comparisons can be made, the evaluator must also know when the information contained in the database was collected. This is critical because older datasets may not be as relevant to or representative of current conditions and climate that may impact sample participants. For example, school performance data collected before the implementation of the U.S. *No Child Left Behind* legislation may not reflect some of the changes in assessment of student learning as a result of the legislation. Additionally, selecting a database containing the most current data and information will help avoid or minimize the effect of historical artifacts.

In sum, each potential public database should be closely examined to minimize or account for any major validity threats such as selection bias, history, and maturation (Shadish et al., 2002). These validity threats are of particular concern when using a database-generated matched sample. In the next section, we offer an example from our own work of how public datasets can be used in evaluation studies. Further discussion of validity threats follows the example presented.

DATABASE USE EXAMPLE

The program being evaluated was an after-school tutoring program that we shall call A.S. Tutoring (a pseudonym). A.S. Tutoring served approximately 150 students, with most program participants residing in affordable housing developments. The program was delivered onsite at the various housing development locations. Thus, the program provided students with an opportunity to do their homework, receive individualized tutoring in verbal and math skills, and acquire computer skills close to their homes. The two primary goals of the program were to improve academic achievement and to promote pro-social behaviours.

As is the case with most smaller educational interventions delivered to students free of charge by nonprofit organizations, the evaluation budget was very tight (US$5,000). However, program staff were anxious to understand the program's impact and to have data to explore some of their "hunches" about the program's worth. We were referred to the group, and as we discussed the possibilities for the evaluation it quickly became apparent that, like many of the programs we had encountered in previous work, there were several context-specific constraints. These constraints included but were not limited to minimal funds for evaluation, a program that was well underway and struggling to meet its goal of serving a particular number of students, and limited time to complete the evaluation. Nevertheless, an evaluation of the program was needed to examine the program's two primary goals. Given the restraints imposed on the evaluation, we proposed using two separate public datasets to examine the impact of A.S. Tutoring on program participants' academic performance and pro-social behaviour.

To obtain an understanding of the program's impact on students' academic achievement, we compared participants' standardized test scores relative to others in the state using data from the California Department of Education (CDE) California STAR exam. We selected the CDE exam, as opposed to another standardized achievement test, for our comparison measure for two reasons. First, A.S. Tutoring could obtain CDE STAR exam scores for each program participant from their local public schools. Second, the California Department of Education website publicly shares average standardized test scores for all schools in California. An important feature of this database is that it allows users to select average standardized scores for each school disaggregated by grade level, ethnicity, gender, and the year the exam was given. Thus, with data on each participant's ethnicity, gender, school name, year(s) in which the standardized exam was taken, and students' test scores we could construct an appropriate comparison group from the CDE database. To illustrate, we collected the following information for A.S. Tutoring student participant X: Latino, female, attended Roosevelt Elementary, in fifth grade, took the exam in 2003, and scored a 320 on the math component.

Using the information from the CDE website, the evaluation team was able to compare student X's performance to 2003 average standardized math scores for all students who attended Roosevelt Elementary, were Latino, female, and in the fifth grade.

The comparison offered an examination of how each program recipient was performing relative to their peers at their particular school. This information provided valuable insights and a more nuanced and precise understanding of the program's impact on students' academic performance. To illustrate, the analysis of the average standardized scores of only A.S. Tutoring students revealed that most students were performing below proficiency (according to CDE guidelines). But when comparing A.S. Tutoring student scores to the scores of the matched group derived using the CDE database, it was revealed that A.S. Tutoring students were outperforming their peers. It should be noted that in the absence of random sampling no casual inferences could be made; however, important information about program impact was gained from our comparisons.

The second program goal, improving pro-social behaviour, was defined by the program directors and staff as improving students' behaviour in school, improving their relationships with parents, and improving their behaviour in their community. After searching available public databases, the evaluation team identified the National Longitudinal Survey of Adolescent Health (ADD Health) database. The ADD Health database was selected for use in this particular study because it contains a nationally representative sample of students from different communities and socioeconomic status with broad eligibility criteria (Table 2). In addition, the ADD Health survey contained over 100 variables, which were clustered in sections. Sections covered a wide array of information from measures of self-esteem to measures of substance use. Of specific concern for our evaluation were items related to student behaviour in school and students' perceptions of their relationships with parents, friends, and others. The items included in the ADD Health study were relevant to our study and so we determined that this dataset was indeed appropriate for our purposes. A weakness of the ADD Health data for our study was the timing of data collection: ADD Health collects data in waves every three to five years, with the most current year (2001) occurring about one and one-half years prior to our evaluation of the A.S. Tutoring program.

Table 2
National Longitudinal Survey of Adolescent Health (ADD Health) Information

Who
- Adolescents, in Grades 7–12
- Stratified, random sample of all high schools in the U.S.
- A school was eligible for the sample if it included an 11th grade and had a minimum enrollment of 30 students

When
- Three waves in 1994, 1996, and 2001

Why
- ADD Health seeks to examine how social contexts (families, friends, peers, schools, neighbourhoods, and communities) influence adolescents' health and risk behaviours

How
- In-school questionnaire
- In-home interviews

To ensure that A.S. Tutoring program staff agreed with our assessment that ADD Health would be an appropriate comparison dataset, a copy of the ADD Health survey was presented to program directors and staff for review. Discussions between the program staff and the evaluation team ensued, and a set of items were selected for use in our study based on their relevance to the intended A.S. Tutoring program goals. An additional set of demographic items was selected so that A.S. Tutoring program participants' responses could be compared to an appropriately matched group drawn from the ADD Health sample. All ADD Health items were included in a survey administered to Grades 7–12 A.S. Tutoring participants.

Students were matched on ethnicity, socioeconomic status, grade level, school size, region (west coast), and living environment (rural, urban, and suburban). Student responses were compared to the matched sample and descriptively presented. This comparison provided a measure of how A.S. Tutoring recipients' pro-social behaviour compared to a sample of students in the same age group, economic status, and living conditions. The analysis also revealed important information for staff about the program's impact on participants. For example, A.S. Tutoring students typically felt safer in their neighbourhood and community than did the ADD Health sample students. This was relevant because an aim of the program was to create a safe and supportive environment for students. In addition to the good news, our comparisons helped to identify areas where A.S. Tutoring students were not doing as well as the ADD Health sample. This information was used as a preliminary indicator or warning sign for program staff and directors, and provided them with guidance on where they should focus attention and resources. Again, no casual inferences could be made, but the comparisons between A.S. Tutoring students and ADD Health-matched students provided useful information about how program recipients behaved relative to others.

DISCUSSION

Creating comparison groups from public databases to understand program impact is not as decisive as using randomized control trial design. Without the random sampling, any inferences made would suffer from validity threats (Shadish et al., 2002). The obtained results could be attributed to historical artifacts, differing maturation rates of participants, and differing study drop-out rates (Shadish et al., 2002). For example, differences in responses between A.S. Tutoring students and the ADD Health students could be due to selection bias. It may be the case that students in A.S. Tutoring had parents that were more interested in their child's academic welfare than were the parents of students in the ADD Health comparison group, as a result of which the program parents enrolled their children in the after-school tutoring program. This in turn may have affected the way students viewed school and their relationship with their parents. This possible interpretation cannot be ruled out using the matched database group design. Again, the evaluator must be aware of these threats to validity and must try to minimize these threats whenever possible.

This leads to the question: Considering these possible threats to validity, why should the evaluator use databases to create a matched group sample? If there is no way to find a viable comparison group, then this approach offers a relatively feasible alternative that provides insight into the relative performance of program participants. The evaluator must still acknowledge that "The risks of error implicit in archival [databases] sources are not trivial, but … if they are recognized and accounted for by multiple measurements techniques, the errors need not preclude the data" (Webb et al., 1966, p. 53). Even with the validity risks, this approach provides the evaluator with some relative performance measures that can be used to inform program stakeholders. In addition, this is a relatively inexpensive way of answering the "compared to what?" question and can be addressed in a relatively short period of time. By providing evaluation stakeholders with a viable comparison group, the evaluation can highlight areas of success and areas in need of further attention and improvement. The major drawback is the rigidity of the database structure, which can limit the kinds of evaluation questions that can be answered, and dictate the construct definitions of outcome measures. These advantages and drawbacks should be shared and discussed with program stakeholders to ensure a common understanding of the validity of findings and responsible decision making.

NOTES

1. The defining feature of an RCT design is the random assignment of participants to a control and an experimental group where quasi-experiments do not have randomization as a feature of the study design. Thus, a primary distinction between RCTs and quasi-experiments is random assignment.

2. It should be noted that databases such as NCES and ADD Health require an application process describing how the data will be used, and the evaluator has to wait until the application is approved to gain access to the data.

REFERENCES

Bertrand, W., Mock, N.B., & Franklin, R. (1981). Social indicators: Their use in evaluation research. In R. Conner (Ed.), *Methodological advances in evaluation research*. Beverly Hills, CA: Sage.

Kirk, R. (1995). *Experimental design: Procedures for the behavioral sciences*. New York: Brooks/Cole.

National Research Council. (2002). *Scientific research in education. Committee on scientific principles for education research*. Washington, DC: National Academy Press.

Patton, M.Q. (1997). *Utilization-focused evaluation: A new century text* (3rd ed.). Thousand Oaks, CA: Sage.

Scriven, M. (2004). *Claremont Graduate University debate on establishing causality*. Retrieved October 23, 2005, from <http://www.cgu.edu/sbos/pdw.html#Debate>.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs: For generalized causal inference*. New York: Houghton Mifflin.

Stewart, D. (1984). *Secondary research: Information sources and methods*. Beverly Hills, CA: Sage.

Webb, E., Campbell, D., Schwartz, R., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally.

Weiss, C. (1998). *Evaluation: Methods for studying programs and policies* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Tarek Azzam is senior research associate at the Institute of Organizational and Program Evaluation Research at Claremont Graduate University. His research focuses on understanding the impact of context on evaluation practice.

Christina Christie is an associate professor and associate director of the Institute of Organizational and Program Evaluation Research, School of Behavioral and Organizational Sciences, Claremont Graduate University. Her research focuses on understanding the relationship between evaluation theory and practice.