Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 19, Number 6, June 2014 ISSN 1531-7714

Quasi-Experiments in Schools: The Case for Historical Cohort Control Groups

Tamara M. Walser, *University of North Carolina Wilmington, NC*

There is increased emphasis on using experimental and quasi-experimental methods to evaluate educational programs; however, educational evaluators and school leaders are often faced with challenges when implementing such designs in educational settings. Use of a historical cohort control group design provides a viable option for conducting quasi-experiments in school-based outcome evaluation. A cohort is a successive group that goes through some experience together, such as a grade level or a training program. A historical cohort comparison group is a cohort group selected from pre-treatment archival data and matched to a subsequent cohort currently receiving a treatment. Although prone to the same threats to study validity as any quasi-experiment, issues related to selection, history, and maturation can be particularly challenging. However, use of a historical cohort control group can reduce noncomparability of treatment and control conditions through local, focal matching. In addition, a historical cohort control group design can alleviate concerns about denying program access to students in order to form a control group, minimize resource requirements and disruption to school routines, and make use of archival data schools and school districts collect and find meaningful.

The current education research and evaluation climate favoring experimental and quasi-experimental studies has left educational program evaluators and school leaders with the task of implementing quality control group studies in school settings where the feasibility and appropriateness of such designs is often in question. The purpose of this article is to describe the use of a historical cohort control group as a viable option for conducting quasi-experimental outcome evaluations in schools; that is, the selection of a cohort control group from pre-treatment archival data matched to a group of students currently receiving an educational treatment (i.e., intervention).

This article includes background information describing the rationale for using a historical cohort control group, including challenges to implementing experimental and quasi-experimental designs in school-based evaluation studies; a description of historical cohort control group designs; and notable threats to the validity of study findings and other considerations when using a historical cohort control group design. The audience for this article is educational program evaluators and school leaders.

Background

The rationale for using a historical cohort control group is based on the emphasis placed on using social science methods that address causal questions of program effectiveness, the related need for educational program evaluation credibility in the current accountability climate, and the need to conduct evaluations that are feasible and appropriate in the real world of school-based outcome evaluation. Under the Elementary and Secondary Education Act (1965), as reauthorized under the No Child Left Behind Act of 2001 (NCLB), *scientifically based research*, one of the four pillars of NCLB, is defined as experimental or quasi-experimental studies (U.S. Department of Education, 2005). Although NCLB is often associated with student testing, terms such as "evidence-based decisions" and "scientifically-based research" appear 111 times in the pages of the NCLB legislation (Wilde, 2004).

In addition to the call for scientifically based research under NCLB, the U.S. Department of Education's Office of Research and Improvement was reauthorized in 2002 with the Education Sciences Reform Act, and became the Institute of Education

Sciences. With this change came new parameters for the type of research funded by IES, emphasizing empirical methods using observation or experiment (Eisenhart & Towne, 2003). Thus, the U.S. Department of Education has enacted a "priority" in grant competitions favoring experimental studies; or when random assignment is not possible, quasi-experimental studies that include matched control groups, regression-discontinuity designs, or single-subject designs (Mills, 2008).

Further, the U.S. Department of Education's What Works Clearinghouse, which also began in 2002, emphasizes studies of causal effectiveness using student achievement test data (Eisenhart & Towne, 2003). Acceptable designs for the What Works Clearinghouse include high quality experiments, quasi-experiments, regression-discontinuity designs, and single subject research studies (U.S. Department of Education, 2008). Finally, the U.S. Department of Education's Institute of Education Sciences and the National Science Foundation jointly published "Common Guidelines for Education Research and Development" in which impact studies—efficacy, effectiveness, and scale-up studies—require experimental or quasi-experimental designs (U.S. Department of Education & National Science Foundation, 2013).

What does this mean for educational evaluators and school leaders?

NCLB requires that educators use instructional programs and methods "proven" to be effective. Scientifically-based research must be used for program development to ensure a scientific basis for the program; and for program evaluation to determine program effectiveness (Slavin, 2003). For states, school districts, and schools; this means that instructional materials and methods adopted and used must have research-based evidence of effectiveness. Further, locally developed and/or implemented programs are held to the same standard (Schmitt & Whitsett, 2008). This also means that when conducting federally-funded outcome evaluations, often referred to as effectiveness or impact studies, NCLB guidelines favor experimental controls when possible (Mills, 2008) and the use of student achievement test data (Eisenhart & Towne, 2003).

What are some challenges to implementing control group studies in schools?

Control group studies may be experimental or quasi-experimental, with the latter being more common in educational program evaluation. Although experimental studies requiring random assignment of participants to either treatment or control conditions are important for responding to causal questions of program effectiveness, such studies have not been common in education (Boruch, 2007; Cook, 2003). Reasons commonly cited for the lack of use include not wanting to deny program access to some students for the sake of forming a control group and the resource requirements for implementing experiments (Baruch, 2007; Baughman, 2008), as well as concerns about disruption to school routines, and the lack of value placed on experiments by educational program evaluators (Cook, 2003).

Quasi-experimental studies are similar to experiments in that they compare treatment to nontreatment conditions; however, participants are not randomly assigned to conditions. Instead, evaluators use a nonequivalent control group for comparison (Cook, 2003). Common quasi-experimental control group designs include the use of intact treatment groups (e.g., classrooms, schools) matched to control groups on demographic and other key variables. The lack of random assignment increases threats to the internal validity of study results—the ability to attribute study results to the treatment and not some other source or sources. In particular, although students, classrooms, and/or schools may be matched on certain known and observable variables, they may differ on other unknown and/or unobservable variables in ways that differentially impact results. In addition, as with experiments, questions of the feasibility and appropriateness of quasi-experimental studies in schools present challenges to their use.

Using a Historical Cohort Control Group

A cohort is a group of people who have similar demographic or statistical characteristics (dictionary.com, n.d.). In education, the term is often used to identify successive groups that go through a grade level, an educational program, or a training program. According to Shadish, Cook, and Campbell (2002):

> [C]ohorts are particularly useful as control groups if (1) one cohort experiences a given intervention and earlier or later cohorts do not; (2) cohorts differ in only minor ways from their contiguous cohorts; (3) organizations insist that an intervention be given to everybody, thus precluding simultaneous controls and making possible only historical controls; and (4) an organization's archival records can be used for

> constructing and then comparing cohorts (pp. 148-149).

If an earlier cohort does not receive a given treatment, they can serve as a *historical cohort control group* for a current group that is receiving the treatment. Thus, it is possible to conduct a quasi-experiment comparing the outcomes of a treatment group that is currently receiving a treatment to those of a historical cohort control group that did not receive the treatment. In research, the term cohort is also used to refer to any group that is repeatedly measured over time, as in a longitudinal or panel study; however, this is a different use of the term (Shadish et al. 2002).

Using a historical cohort control group is a viable option due to its feasibility and appropriateness in school settings, where, as mentioned previously, there are often challenges to implementing experimental or commonly used quasi-experimental designs. Students who are eligible for a program are not denied access so that a control group can be formed; and because control group data come from archival data sources, no new data collection is needed, decreasing the resources required to conduct the evaluation, as well as disruption to school routines. Due to NCLB, more data are available (Guillen-Woods, Kaiser, & Harrington, 2008), including student achievement test data favored under NCLB (Eisenhart & Towne, 2003) andlongitudinal data (Azin & Resendez, 2008).

Finally, the goal of matching a control group to a treatment group is to achieve as much overlap of the two groups, or distributions, as possible in order to minimize initial nonequivalence. T. D. Cook and W. R. Shadish (personal communication, August 7, 2008) have suggested using a *local, focal, nonequivalent intact control group* to maximize overlap; that is, using a local group from the same site (e.g., school, school district), and matching on focal indicators that are related to the outcome. A historical cohort control group is an example of this.

An example of a historical cohort control group design used in a school-based evaluation study is Stockard's (2011) evaluation of a direct instruction reading program in primary grades in three rural school districts. Using existing student characteristics data, Dynamic Indicators of Basic Early Literacy Skills (DIBELS) assessment data, fourth grade state assessment data, and implementation data; Stockard was able to conduct several analyses comparing district, state, and national data. Results indicated that students in cohorts that received full exposure to the program, those who began the program in kindergarten and whose teachers implemented the program with fidelity, had higher scores than students in cohorts that received less exposure.

Al-Iryani, Basaleem, Al-Sakkaf, Crutzen, Kok, and Bart van den Borne (2011) also used a historical cohort control group, in addition to a concurrent control group, in their evaluation study of a school-based peer intervention for HIV prevention among students. A survey to determine knowledge and attitudes related to HIV was administered concurrently to students who received the peer intervention and students who did not receive the intervention. In addition, a historical cohort control group was randomly selected from a sample of students from the same schools who completed the survey in 2005. Thus, survey results for the historical cohort control group were compared to those of the intervention group, in addition to comparisons between the intervention group and concurrent control group. Based on results, the authors concluded that the intervention improved knowledge on modes of transmission and prevention and decreased levels of stigma and discrimination.

Validity and Historical Cohort Control Group Designs

When conducting a quasi-experiment using a historical cohort control group, evaluators are faced with the same threats to the validity of study findings as they would be when implementing any quasi-experiment. Early conceptualizations of validity characterized it as covering two aspects, internal and external validity. Internal validity was generally defined as the ability to attribute observed differences in outcomes to the treatment under investigation. External validity was defined as the ability to generalize study outcomes to different contexts. Within each of these categories of validity were several specific validity threats—i.e., issues that could decrease validity of study findings.

In their often-cited book on experimental and quasi-experimental design, Campbell and Stanley (1963) identify 12 common threats to internal and external validity and describe the level of these threats given different research design options. More recent conceptualizations of validity have expanded these categories of validity to include:

- Statistical conclusion validity, which focuses on the validity of inferences given potential analysis issues (e.g., low statistical power, restriction of range).

- Internal validity, which focuses on the validity of inferences given potential issues that can influence outcomes instead of or in addition to a causal relationship between treatment and outcome (e.g., selection, history).
- Construct validity, which focuses on the validity of inferences given potential issues related to the constructs that represent the specifics of the study (e.g., construct confounding, novelty and disruption effects).
- External validity, which focuses on the validity of inferences given potential issues with the ability to generalize a causal relationship from one context (setting, persons) to another (e.g., interaction of the causal relationship with units, context-dependent mediation) (Shadish et al., 2002).

The following sections include a discussion of specific validity threats that are particularly relevant when using a historical cohort control group. As previously mentioned, a historical cohort control group design is prone to the same validity threats as any quasi-experimental design; thus, only those threats that educational evaluators and school leaders should be more cognizant of when using a historical cohort control group design are discussed. These notable threats fall into the categories of internal validity and construct validity. For a complete description of the 37 identified threats across the 4 categories of validity, see Shadish et al., 2002.

Threats to Internal Validity

Internal validity threats that are notable when using a historical cohort control group design include selection, history, maturation, regression, testing, and instrumentation. Each threat is discussed in the following sections.

Selection threats occur when observed differences may be due to differences in participants that existed prior to the study (Shadish et al., 2002). Selection poses a considerable threat to any quasi-experimental study, because there is no random assignment. Use of a historical cohort control group design has the potential to maximize overlap of treatment and control groups; that is, to minimize selection differences between groups that would pose threats to internal validity. Generally, cohorts are considered more similar to each other than most nonequivalent (nonrandomized) comparison groups: "The crucial assumption with cohorts is that selection differences are smaller between cohorts than would be the case between noncohort comparison groups" (Shadish et al., 2002, p. 149).

On the other hand, selection threats may increase depending on when data were collected from the historical cohort. For example, in a school-based evaluation study, students in treatment and control cohort groups may have attended or currently attend the same school and may be from the same community, but depending on the time period in which the historical cohort control group data were collected, there could be differences in treatment and control cohort characteristics due to changes in the demographics of the school and community. Thus, it is important to understand the context within which a historical cohort control group design is used, to document changes in cohort characteristics that have occurred over time, and to consider a different design option if those changes greatly increase selection threats.

History threats are problematic when other events that occurred during the time of treatment could have caused or impacted observed differences (Shadish et al., 2002). History threats are an issue when using a historical cohort control group design due to differences in the time of data collection. With concurrent treatment and control conditions, it is possible that any influences of history on outcomes would occur in both conditions. As stated in Shadish et al. (2002), "Only when a nonequivalent control group is added to the design and measured at exactly the same time points as the treatment cohorts can we hope to address history" (p. 151). Thus, the use of a historical cohort control group cannot address history threats.

Maturation threats are a concern when "naturally occurring changes over time could be confused with a treatment effect" (Shadish et al., 2002, p. 55). This is a common concern in studies of children who are developing cognitive skills as part of a normal progression. For example, first grade student gains in reading comprehension may be due to natural development instead of or in addition to a particular reading treatment. Maturation threats may be reduced by using a historical cohort control group design. Because cohort groups are, by definition, similar in characteristics and from the same location (e.g., cohorts in school-based evaluation studies would be the same age and/or grade level), changes that occur in the treatment cohort group that could be attributed to maturation should have similarly occurred in the historical cohort control group; thus, any observed differences between the groups would not likely be the result of maturation. When using a historical cohort control group design, it is important that there is as much overlap in characteristics between the control cohort and treatment cohort as possible. Using local, focal matching, where cohorts are from the same location and are matched on characteristics focal to the study, supports this overlap (T. D. Cook & W. R. Shadish, personal communication, August 7, 2008).

Regression threats occur when participants are selected for a study due to extreme scores. They will often have less extreme scores on subsequent measures and this shift in scores can be confused with a treatment effect (Shadish et al., 2002). In other words, if study participants are selected due to a deficit in some area, such as reading ability, subsequent outcomes related to reading ability may be higher due to a phenomenon known as "regression to the mean." Regression threats may be less of a concern when using a historical cohort control group design, because cohorts are characterized as having similar characteristics and would be chosen as study participants based on local, focal criteria. Any observed differences due to regression would occur in the historical cohort control group and the treatment cohort group.

Instrumentation threats occur when a measure changes over time, impacting study outcomes. In addition to actual changes in the measure, changes in the conditions within which the measure is administered can also impact observed differences. Instrumentation can be particularly problematic when using a historical cohort control group design, due to the time interval between data collection for the historical cohort control group and the treatment cohort. For example, those who administered measures for the historical cohort control group may be different than those who administer measures for the treatment cohort, which could result in differences in administration that impact scores. Further, even if the same people administer the measures for both groups, their skills in administration may improve or otherwise change over time. Thus, when using a historical cohort control group design, it is important to understand and document the test conditions for cohorts, noting discrepancies that may introduce instrumentation threats.

Threats to Construct Validity

Construct validity threats that are notable when using a historical cohort control group design include reactivity to the experimental situation, compensatory equalization, compensatory rivalry, resentful demoralization, and treatment diffusion. Each threat is discussed in the following sections.

Reactivity to the experimental situation occurs when participants' perceptions related to being in a study impact their behaviors and influence study outcomes (Shadish et al., 2002). For example, if a student knows she is in a study, she may work harder to do well on tests. In general, use of a historical cohort control group design can decrease such reactivity issues, because measures used for the historical cohort control group and treatment cohorts are measures that are routinely used as part of monitoring and evaluation. In the case of school-based evaluation studies, measures could include progress monitoring, benchmark, and annual assessments used by the school district to monitor student achievement. Thus, they would be seen by students as "business as usual."

Compensatory equalization occurs when the treatment includes desired resources or services not afforded to the control group participants, resulting in actions to provide compensatory resources or services to the control group by administrators or staff; thus, confounding study outcomes (Shadish et al., 2002). Using a historical cohort control group design can eliminate compensatory equalization threats, because the historical cohort and treatment cohorts are not concurrent. Thus, no compensatory resources or services would be provided to the historical cohort control group, nor would any be taken away from the treatment group in an effort to equalize.

Compensatory rivalry occurs when control group participants are motivated to show that they can do as well as those in the treatment group (Shadish et al., 2002). This could include, for example, control group teachers in a school-based evaluation study being more motivated to ensure their students demonstrate outcomes associated with the study, as well as the students themselves "competing" with students in the treatment group. Similar to compensatory equalization threats, compensatory rivalry threats are not an issue when using a historical cohort control group design. Because the control group is "historical" and only archival data from the group are used, historical cohort control group participants would not know they did not receive a given treatment.

Resentful demoralization is the flip side of compensatory rivalry and occurs when control group participants are so demoralized by not being selected to receive the treatment that this negatively impacts their outcomes; thus, any observed differences in treatment and control conditions are influenced by these attitudes (Shadish et al., 2002). As with compensatory equalization and compensatory rivalry, use of a historical cohort control group can eliminate resentful demoralization threats.

Treatment diffusion occurs when the control group receives some or all of a treatment (Shadish et al., 2002). For example, in a school-based evaluation study, the same teacher may provide treatment and control group instruction; thus, the teacher may use some of the treatment strategies with the control group. Even if the treatment and control groups have different teachers, the control group teacher may hear about some of the treatment strategies and implement them. Of course, similar to the other construct validity threats previously discussed, treatment diffusion is not an issue when using a historical cohort control group, because treatment and control conditions are not concurrent and the control condition occurs prior to introduction and implementation of the treatment.

Additional Considerations

Perhaps the greatest benefit of using a historical cohort control group in terms of increasing the validity of study findings is the potential to maximize overlap of treatment and control groups; that is, to minimize selection differences between groups that would pose threats to internal validity. Generally, cohorts are considered more similar to each other than most nonequivalent (nonrandomized) comparison groups: "The crucial assumption with cohorts is that selection differences are smaller between cohorts than would be the case between noncohort comparison groups" (Shadish et al., 2002, p. 149). Regardless, it is critical for the evaluator to investigate validity threats—to understand study context and document and report issues with validity.

Another consideration is to think of research design as a layering process as opposed to a process of choosing one "textbook" design. Shadish et al. (2002) discuss adding design elements to strengthen validity and offer related suggestions throughout their text. Thus, using a historical cohort control group may be one element of a larger research design that could also include, for example, concurrent control groups. Another option to strengthen validity is the use of multiple historical cohort control groups; that is, using archival data from historical cohort control groups for as many years back as possible. Similarly, archival data for the treatment cohort could be used as multiple pretests.

Finally, standards for quality evaluation require more than addressing study validity. *The Program Evaluation Standards* (Yarbrough, Shulha, Hopson, & Caruthers, 2011) provide guidance on the design, conduct, and evaluation of evaluations. Thirty standards are categorized as utility, propriety, feasibility, accuracy, and accountability standards. Thus, these standards should be considered in addition to addressing study validity.

Conclusion

Using a historical cohort control group is a viable option for addressing causal questions of effectiveness in school-based outcome evaluation. Given the push for control group studies, as well as common challenges to implementing such studies in school settings, using a historical cohort control group (a) alleviates concerns about denying program access to students in order to form a control group, (b) minimizes resource requirements and disruption to school routines, (c) makes use of the archival data schools and school districts collect and find meaningful, and (d) can reduce initial noncomparability of treatment and control groups through local, focal matching. Although a historical cohort control group design is prone to the same threats to validity as any quasi-experimental study, use of the design may decrease threats such as selection, reactivity, compensatory equalization, compensatory rivalry, resentful demoralization, and treatment diffusion. In addition, using a historical cohort control group as a design element as part of a larger study can strengthen validity. Finally, for program evaluation, it is important to also use *The Program Evaluation Standards* (Yarbrough et al., 2011) in the design and conduct of evaluations.

References

Al-Iryani, B., Basaleem, H., Al-Sakkaf, K., Crutzen, R., Kok, G., & van den Borne, B. (2011). Evaluation of a school-based HIV prevention intervention among Yemeni adolescents. *BMC Public Health, 11*(279), 1-10.

Azin, M., & Resendez, M. G. (2008). Measuring student progress: Changes and challenges under No Child Left Behind. In T. Berry & R. M. Eddy (Eds.), *Consequences of No Child Left Behind for educational evaluation. New Directions for Evaluation, 117*, 71-84.

Boruch, R. (2007). Encouraging the flight of error: Ethical standards, evidence standards, and randomized trials. In G. Julnes & D. J. Rog (Eds.), *Informing federal policies on evaluation methodology: Building the evidence base for method choice in government sponsored evaluation. New Directions for Evaluation, 113*, 55-73.

Baughman, M. (2008). The influence of scientific research and evaluation on publishing educational curriculum. In T. Berry & R. M. Eddy (Eds.), *Consequences of No Child Left Behind for educational evaluation. New Directions for Evaluation, 117*, 85-94.

Cook, T. D. (2003). Why have educational evaluators chosen not to do randomized experiments? *The ANNALS of the American Academy of Political and Social Science, 589*(1), 114-149.

Eisenhart, M., & Towne, L. (2003). Contestation and change in national policy on "scientifically based" education research. *Educational Researcher, 32*, 31-38.

Guillen-Woods, B. F., Kaiser, M. A., & Harrington, M. J. (2008). Responding to accountability requirements while promoting program improvement. In T. Berry & R. M. Eddy (Eds.), *Consequences of No Child Left Behind for educational evaluation. New Directions for Evaluation, 117*, 59-70.

Mills, J. I. (2008). A legislative overview of No Child Left Behind. In T. Berry & R. M. Eddy (Eds.), *Consequences of No Child Left Behind for educational evaluation. New Directions for Evaluation, 117*, 9-20.

Schmitt, L. N. T., & Whitsett, M. D. (2008). Using evaluation data to strike a balance between stakeholders and accountability systems. In T. Berry & R. M. Eddy (Eds.), *Consequences of No Child Left Behind for educational evaluation. New Directions for Evaluation, 117*, 47-58.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference (2nd ed.).* Boston: Houghton Mifflin Company.

Slavin, R. E. (2003). A reader's guide to scientifically-based research: Learning how to assess the validity of education research is vital for creating effective, sustained reform. *Educational Researcher, 60*(5), 12-16.

Stockard, J. (2011). Increasing reading skills in rural areas: An analysis of three school districts. *Journal of Research in Rural Education, 26*(8), 1-19.

U.S. Department of Education. (2005). Scientifically-based evaluation methods (RIN 1890-ZA00). *Federal Register, 70*(15), 3586-3589.

U.S. Department of Education. (2008, December). *What Works Clearinghouse standards and procedures document (Version 2.0).* Washington, DC: U.S. Department of Education, Institute of Education Sciences. Retrieved June 14, 2009 from http://ies.ed.gov/ncee/wwc/

Wilde, J. (2004, January). *Definitions for the No Child Left Behind Act of 2001: Scientifically-based research.* Washington, DC: National Clearinghouse for English Language Acquisition and Language Instruction Education Programs, The George Washington University. Retrieved April 10, 2008 from http://www.ncela.gwu.edu/resabout/Research_definitions.pdf

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users (3rd ed.).* Thousand Oaks, CA: Sage.

Citation:

Walser, Tamara M. (2014). Quasi-Experiments in Schools: The Case for Historical Cohort Control Groups. *Practical Assessment, Research & Evaluation*, 19(6). Available online: http://pareonline.net/getvn.asp?v=19&n=6

Author:

Tamara M. Walser
Associate Professor and Director of Assessment and Evaluation
Watson College of Education
University of North Carolina Wilmington
601 S. College Road
Wilmington, NC 28403
Email: walsert [at] uncw.edu