

Experimental Design Considerations

Key Factors to Consider in the Design and Implementation of Experimental Design Evaluations of Teaching American History Projects

Patricia A. Muller, Ph.D

1. Determine if you have an adequate sample size

Determining adequate sample size is complex, and is dependant upon a variety of factors such as the strength of the program effect, the sensitivity and alignment of the instrument being used to measure the outcome, etc. Although there is no simple “right answer,” a minimum sample size of approximately 60 teachers (plus their classes) – 30 in the TAH program group and 30 in the control group - is generally needed to produce strong evidence about a project’s effect such as TAH on student achievement¹.

Although it is possible that impact might be detected with smaller sample sizes if the true effect is large, this minimum sample of 60 teachers will allow the evaluation to identify TAH programs that have a modest or large effect on student achievement, not just those with a large effect; and this sample size increases the evaluation’s ability to also examine impact on teacher content knowledge, as well as student achievement.²

As noted above, however, sample size estimations are not simple. For example, these estimates assume that the TAH program provides roughly the same professional development to all participating teachers. However, many TAH programs provide either different professional development to different teachers (e.g. one summer course for

-
1. This sample size estimate is based on the following assumptions: the desired power for the study is .80; the project’s true effect size is modest in size (e.g. at least .2 standard deviations); each teacher has 35 students total in his or her American History classes; the intra-class correlation is 0.075; a covariate (baseline test scores) with a .8 correlation with outcomes is used on estimating the project’s effect; the study seeks to estimate the project’s effect at the .05 level of significance in a two-tailed test; and the study obtains outcome data for 80% of the original sample of teachers and students (Coalition for Evidence-Based Policy, 2005).
 2. Measuring impact on teacher outcomes such as teacher content knowledge generally requires a larger sample size than that needed to measure the effect on student achievement because the statistical power from including teachers with their classes of students is lost.

middle school teachers, another for high school teachers), or provide options for teachers to choose from a variety of different TAH professional development programs in a manner that allows for varying degrees of involvement and participation (e.g. one teacher might attend the summer course, and three additional sessions during the school year while another teacher attends the same summer course but only one additional session). In these instances where teachers receive varying degrees or types of TAH professional development, then the minimum sample size increases.

2. Ensure the Integrity of the Random Assignment

Given that random assignment is integral to experimental design evaluation, the project administrators and evaluator need to be committed to ensuring the integrity of the random assignment process, including adhering to the following:

1. Institutional independence of the assignment process. The random assignment process should be conducted by a third party researcher independent of the TAH project and evaluator(s).
2. Centralized procedures and decision-making. To ensure that the random assignment process is not undermined, all decisions related to departures from the established process should be made by the objective third party researcher.
3. Quality control system. After assignment is completed, the evaluator(s) should audit the randomization process to ensure its fidelity and detect any flaws. All departures from random assignment should be tracked and reviewed to minimize the chances for subversion of the random assignment process.
4. Analysis of group equivalency. In smaller trials it is possible for random assignment to produce, by chance, intervention and control groups that differ systematically in various characteristics. Therefore, the evaluator(s) should examine any available pre-program data (e.g. teacher characteristics, classroom-level student test scores, etc.) to ensure pre-intervention group equivalence.

In addition, the students should be assigned to their History classes through the schools' usual scheduling process, rather than based on which teachers are participating in the TAH program. Assignment of students based on which teachers are in the TAH program would undermine the randomization.

3. Select Appropriate Outcome Measures

Unfortunately, unlike content areas such as mathematics and language arts, most schools do not already administer achievement tests that can be used for the experimental evaluation of TAH programs. Therefore, the evaluation will likely need to include the selection and administration of an instrument to measure outcomes.

The TAH project goals need to guide the selection of outcome measures for the evaluation. For example, if the primary goal of a TAH program is to increase student knowledge of the state's history, then the selected outcome measure needs to focus specifically on the targeted content knowledge related to the state's history; and if a TAH program's focus is American History during the Civil War era, then the selected outcome measure needs to specifically focus on the Civil War era. Similarly, if a TAH program includes goals related to student motivation and/or interest in American History there should be student outcome measures related to student motivation; and if a primary goal of a TAH



program is to change students' ways of thinking about history or increasing student's historical thinking/analysis skills, then the student outcome measures should also include measures of students' understanding and use of historical thinking/analysis skills.

This close alignment of the selected outcome measure and the primary purpose of the TAH program is critical to the success of the evaluation. If there is not close alignment, then the experimental design evaluation is not likely to detect impact even if there is a significant impact. For example, an experimental design evaluation of a TAH program that has a primary goal of increasing student knowledge of American History during the Civil War era is not likely to detect even large program effects if a general social studies achievement test is used as the outcome measure. The general social studies achievement test might include (at best) one or two items that pertain to American History during the Civil War era, and therefore will not be a reliable or meaningful measure for this particular program.

In addition to alignment issues, the reliability (i.e. consistency, accuracy and reproducibility of a measure) and validity (i.e. appropriateness, meaningfulness and usefulness of the specific inferences made from test scores) of outcome measures should also be taken into consideration. Many instruments have been subjected to psychometric tests that provide indicators of the instrument's validity and reliability.

Given that impact on students is a key TAH program goal, student level measures are generally important to include in experimental design evaluations of TAH. An additional option to consider is including teacher-level outcome measures. Given that it often takes many years for impact on students to become evident, and improving teacher knowledge/skills is a key intermediate goal of TAH, including measures of teacher content knowledge or teaching methodology may provide preliminary evidence of program impact at an early date in time. Unfortunately, as noted above, sample sizes of teachers may not be large enough to generate strong evidence about the effect on teacher content knowledge.

If possible and appropriate for a given TAH evaluation, the quality and rigor of the experimental design evaluation will also be improved by including measures of the following:

- a. “Pre-test” or “pre-intervention” measures. Although the random assignment process is designed to minimize any pre-program differences between the treatment and control groups, it is also helpful to have a “pre-test” or “pre-intervention” measure to ensure the equivalency of the groups, particularly with relatively small sample sizes.
- b. Measures of Project Implementation. If possible, the TAH evaluation should include measures of the extent to which the professional development is implemented in the intended manner, and the extent to which the teachers apply new knowledge and skills learned through the TAH program. These measures of program fidelity and follow-through will help to determine the extent to which any findings of non-significance or impact are due to the program not being implemented in the intended manner rather than the program itself having no effect.



-
-
- c. Measures of “treatment exposure.” Many TAH programs consist of multiple professional development components (e.g. summer sessions, one-day modules during year, mentoring component). Even if a program is not intentionally designed for teachers to participate in only *some* components of the program, there will be at least some teachers that in actuality do not attend all components of the program. For example, including a measure of “treatment exposure” allows the evaluation to differentiate between those teachers who attended a total of 10 hours worth of training from those who attended 80 hours worth of training.
 - d. Long-term educational outcomes. Although it may not be possible in many TAH evaluations due to small sample sizes, attrition or resources, including measures of long-term student outcomes is beneficial for understanding the sustained impact of TAH programs.

4. Take Steps to Minimize Common Threats to Internal Validity

To yield findings that are credible, threats to the internal validity of the study need to be controlled. This can be accomplished by using a strong design that minimizes the potential threats to achieving valid results. The most common threats to internal validity for TAH programs that have not been addressed previously in this document include the following¹:

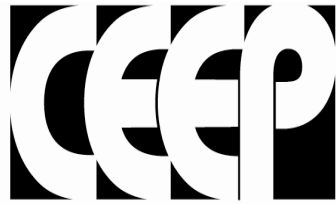
History: Non-TAH events that happen during the treatment may result in effects not caused by the TAH program. For example, teachers may gain knowledge or skills related to American History from outside sources rather than TAH. To guard against such threats to the internal validity of the study, the evaluators should maintain an awareness of other programs occurring in the schools, and collect relevant data related to non-TAH professional development or events that might impact results.

Experimental mortality/attrition: It is not unusual for professional development programs to lose program participants for a variety of reasons (e.g. retirement, teacher mobility, teacher chooses not to participate any longer). Even if the attrition between treatment and control groups is similar, the attrition may be an issue due to the relatively small sample sizes for most of the TAH programs. To maintain the integrity of the randomization, it is important to obtain data for at least 75-80% of the teachers originally randomized, and the students in their classes. Offering incentives for participation for both treatment and control group may help minimize attrition. In addition, collecting and analyzing outcome data for all teachers randomly assigned, even those intervention-group teachers who do not actually complete the TAH program, can help minimize threats to validity. This “intention-to-treat” approach is designed to ensure that the intervention and control groups remain equivalent over the course of the study.

1. Other common threats to internal validity (e.g. selection, maturation) are not addressed here because these particular threats are not generally as pertinent to TAH programs using experimental design evaluation or have been addressed elsewhere in this document.



Produced by:



**CENTER FOR EVALUATION
& EDUCATION POLICY**

On the web at:

www.ceep.indiana.edu

