EDLD704: Methods and Instruments for Data Collection

Dr. Jack Huber

1/23/23

Table of contents

Description of the Course

The purpose of this course is to teach you about research methods and instruments for collecting data. The course will expose you to several research methods and offer guidance for collecting both quantitative and qualitative data. You should finish the course knowing:

- how to select an appropriate research method for investigating a question for a study arising from a problem of practice
- how to conceptualize, develop, and test an instrument for collecting data
- how to evaluate the quality of data collected, and how to express your evidence-based argument in clear, simple prose, using APA format, to a skeptical reader

To provide you opportunity to learn these skills, the course expects you to complete two projects: one quantitative, one qualitative. In each project, you will:

1. frame one or more questions to guide inquiry into a problem of practice
2. select a research method appropriate to the nature and purpose of your inquiry question/s,
3. conceptualize and design a small study using your selected research method,
4. design or select instrumentation to collect data,
5. collect a small sample of data,
6. critically evaluate the quality of your data
7. draw any appropriate inferences or make any appropriate claims from your study.

Questions, Methods, and Data

In your setting, surely you've seen data collected and presented in various ways to get your attention, stimulate your thinking, illuminate an issue, and the like. Maybe you've done such data-related work yourself.

Like appreciation for fine art or food acquired over time, you probably have a sense of "bad" data when you see it.

Do you also have a sense of "good" data when you see it? And can it still be "good" even if you disagree with it?

I want this course to help you collect better data from now on by focusing your attention on *how data are collected* – that is, **methods** – for conducting inquiry, which includes considerations how to collect data in ways that optimize their quality.

The central concerns of this first module are the question and decision of what method to use to carry out a study. As I see it, methods depend, fundamentally, on the nature of the research question: What specific kind of information or understanding does the question seek? Here I make a fundamental and admittedly over-simplistic distinction between **qualities** and **quantities**.

To investigate questions that betray interest in quantities – *"To what extent…?", "How prevalent…?", "What predicts….?"* – one should use quantitative methods. This includes surveys, experiments, quasi-experiments, and the like.

To investigate questions that betray interest in qualities, we employ qualitative designs and instruments. These include interviews, observations, in-depth case studies, analysis of content, and so on.

Often the research question is so self-evident that the choice of method and instrumentation is obvious.

But perhaps just as often we are interested in *both*, and the qualities-quantities distinction is not so clear cut. To illustrate, consider this case study:

A Case Study

In 2006, all public high schools in Washington State were required to administer the state's large-scale assessment, the Washington Assessments of Student Learning (WASL) in mathematics and reading, to all tenth grade students. By law, these students were the first graduating class required to pass the test in order to receive their high school diplomas. High stakes accountability testing was getting "real."

Public educators throughout the state were anxious. Questions abounded:

- How many students would meet the standard? How close are we?
- Which students are less likely to reach the standard?
- What reforms, interventions, or other restructuring are necessary in elementary and middle school to better prepare students for the high school proficiency standard?
- What exactly are the high school proficiency standards in reading and math?
- What reforms, interventions, or other restructuring are necessary in high school to better prepare students who failed the test in tenth grade to pass the test by their senior year?

This state policy was controversial. Many people decried the requirements as fundamentally unfair. Leading psychometricians (experts in test design) criticized the high stakes policies as invalid uses of (largely high quality) standardized tests. Some public educators retired early or found other jobs. Others defended the policies as necessary to bring about long overdue reforms.

At the time, I was somewhere in the middle. It was early in my career in public education. I was employed as a data analyst in my district's curriculum and instruction department, and on the side I was working on my doctorate. My job, in essence, was to help educators understand student achievement data. I was unique because I had come to education not from classroom teaching but the academic world: sociology. I cared what *data* said. Here are the kinds of questions I asked at the time:

- What is the historical and/or social scientific evidence that these high stakes accountability testing policies actually work? Where and when have these policies already worked?
- And what does "actually work" really mean: To improve instruction? To help students overcome demographic disadvantages?
- Part of the theory of action of these accountability policies is "measurement-driven instruction": testing data should provide instructionally valuable feedback. Teachers should look at data; and when they do, they should see and do …. *what*?
- "Data-based decision-making" is all the rage. But what exactly does it mean for a district or school to be "data-driven"? What decisions? What data? Might this look very different from one district or school to another?

- How does a district or school become "data-driven"? By what process of evolution?
- High school teachers already see state assessment at a high level in the summer during inservice days. How often do they look at the state assessment data for their own students? And when they do, how much instructional utility do they derive from the data?
- The policies assume that external accountability pressure will cause teachers to look at data. Can I test that empirically? Do teachers who perceive more pressure tend to look at state assessment data more often than teachers who perceive less pressure?
- Professional learning communities are hailed as a very effective model for organizing and motivating teachers to collaborate. Are high school teachers in professional learning communities more likely to use high school assessment data to improve instruction than those not in professional learning communities?

After 16 years, these questions probably sound dated now. They were on a par with what people were writing at the time and they lended themselves readily to disciplined, systematic inquiry data. More to the point, let's consider the different *kinds* of research questions in this topic.

Questions for quantitative data

The quantitative questions are fairly obvious:

How often do high school teachers use state assessment data? This is a question is a no-brainer because it is about *frequency*, which ranges from *less* frequent ("hardly ever') to *more* frequent ("all the time"). To study this I needed a sample of teachers who varied in their use of data along this range of frequency.

Are teachers who perceive more external accountability pressure to improve test scores *more likely* to examine their own students' state assessment data more often? This too is an unmistakably quantitative research question (or hypothesis). Implied is comparison between two groups (which is "more likely") along scales of intensity for accountability pressure (less to more intense), frequency ("rarely" to "often") of data use. To study this I needed a sample of teachers who varied in their perceptions of accountability pressure and their frequency of data use.

Notice that questions for quantitative data come from an understanding of the situation of enough sophistication to know what the important variables are and how the variables might be related (do the values of one depend on the values of the other). In most cases, quantitative analysis is **deductive**; we know what to look for and we understand the situation well enough to test competing theories or understandings.

Quantitative methods are also appropriate when you want to make generalizations about a population. They seek to show what is **generally true** of a **large number** of "cases" (most often, people).

Questions for qualitative data

Notice the questions that are more clearly about qualities than quantities.

What does it mean for a district or school to be "data-driven"? Nothing here is quantified or quantifiable. The quest is for *attributes* or *states* of "data-driven". The result could be a typology of different kinds of "data-driven"-ness. Or it could be some kind of evolutionary process with beginning and more advanced stages of development.

What does it mean for a teacher to "use" state assessment data? "Using state assessment data" could mean different things among high schools than my understanding from the district office, the professional research literature, and my background in social science. I needed to talk to sample of teachers to ask them to describe in their own words how they use data.

What sense do high school teachers make of state assessment data? Similar to the question above, I needed to ask teachers to describe what (if anything) they learn from state assessment data in their own words.

For each of these questions, the focus is full understanding of a small number of cases (most often, people). Generalization to a large population is *NOT* the point of qualitative methods. Qualitative methods aim to understand what is **deeply true** *of a* **small number** *of cases.*

Your turn

Having considered the different angles for research in this case study, now think about your own problem of practice as it seems to you in your setting or milieu. Maybe this is your nascent capstone project.

Write down your guiding question/s that best capture your true interest.

Then consider the words you've used.

Are you looking to explore something that is not well understood? Do your questions seek understanding of *kinds*, *ways in which*, *processes*, *stages*, *distinctions*, *classes*, *forms*, and the like? Are these things you can average? (No?) Do you seek understanding of the mental *models*, *theories*, *understandings*, and the like, of how someone in your setting of interest perceives something, or understands what they're doing? Do you want their own words? Are you interested in the "theory of action" behind a program or organization? Are you interested in *identities* and *self-understandings*? Are people's own *metaphors* interesting to you? Do you

want to deal primarily with "words" data? If yes, then you may be primarily interested in qualitative methods.

Do you have a good enough understanding of your topic that you know what the *important factors* or *variables* are? Is one variable more important than another? Do you want a sense of *scope*, *estimate*, *size*, *frequency*, *magnitude*, *intensity*, *extent*, *prevalence*, *risk*, *predictability*, *regularity*, or *relationship*? Do you want to deal with primarily with "numbers" and "scale" data? If yes, then you may be primarily interested in quantitative methods.

A final word, for now, about mixed methods:

There are good reasons to use mixed methods. You may want to collect some qualitative data (from interviews, observations) from a few cases to more deeply understand something. With better firsthand understanding you can then develop more accurate survey items, frame more relevant questions and hypotheses, and test competing explanations of something.

My doctoral dissertation was *de facto* mixed methods. It began with qualitative work. From my role in the central office I knew a lot about my topic from a global perspective and from the professional literature, but I did not understand teacher work life very deeply. Interviewing a small sample of them helped me better understand my topic from their perspective. But I didn't stop there; I wanted to make generalizations to a population of teachers. Based on this more sophisticated understanding I was able to frame smarter research questions and better survey items and to specify and estimate more grounded statistical models. My quantitative dissertation study proper owes its quality to the preliminary qualitative work that informed it.

Mixed methods are possible, and may appeal philosophically: "Why choose between the two if I can do both? Wouldn't mixed methods make the most sense and do the most justice to the topic?" True enough. And what more appropriate laboratory for learning different research methods than your doctoral program? But to do any research method well is to negotiate a learning curve, and your time and energy are limited in this fast-paced doctoral program. Do factor that into your discernment of methods. Whatever you decide, I will help you as best I can.

Part I

Quantitative Methods

Surveys

Required reading:

- Chapter 4 (pp. 62-63) in Burkholder et al. (2020)
- Chapter 11 in Burkholder et al. (2020)

Why do a survey?

A survey is an efficient way to collect a large quantity of data on a large number of people in a relatively short amount of time. Then one can use these data to:

- "Explore a topic that has not been previously examined" (Burkholder et al. (2020), p. 163)
- "Explain a relationship between two or more variables of interest" (Burkholder et al. (2020), p. 163)
- "Describe the characteristics or attributes of a population" (Burkholder et al. (2020), p. 163)
- Make generalizations about a population of people
- Get a sense of the scope, extent, magnitude, or prevalence of something
- Measure a construct, such as psychological well-being

Depending on your guiding questions, a survey may be the appropriate method to collect the data you need for your capstone. And at some point in your time in education, you may want or need to conduct a survey. Now is a great time to gain understanding and skill.

Some key terms

Survey methods have their own vocabulary. Following is a list of key terms you should know when reading about and use when undertaking surveys:

Survey	the "method of collecting data from and about people" (Fink, 2009, quoted in Burkholder et al. (2020), p. 161)
Survey instrument	"the tool used to gather data–this term is typically used to differentiate the tool from the survey research it supports" (Burkholder et al. (2020), p. 161)
Questionnaire	"a survey instrument that contains items that the respondent is expected to read and then report his or her own answers" (Burkholder et al. (2020), p. 161)
Form	The body of the survey or test instrument where all of the items are assembled. A survey may use two different forms, such as Form A and Form B, each of which contains the same items in different orders, to examine the effects of item order on responses Item a question on a survey or test that gathers responses from a respondent and creates variation
Response categories	Categories, such as those found on a Likert scale (1=Strongly Agree, 2=Agree, etc.), that a respondent may use to respond to a survey item
Descriptor	A descriptive label, such as "Strongly Agree", that one applies to a response category to make it the response meaningful to the respondent
Respondent	an individual who responds to an item and/or survey instrument
Pilot	a phase of the survey project when an investigator uses instrument to collect sample data for the purpose of improving the instrument and/or data collection procedures
Operational	he final phase of the survey project in which the instrument collects data of sufficient quality to collect "real" data for the purpose of supporting high stakes decisions

Properties of a poor quality survey

We've all seen and/or taken poor quality surveys. Here are a few characteristics of poor quality surveys:

- **The items are too long**. The survey writer is wordy and/or has too much "voice." It's difficult to tell what the respondent is thinking and/or what the respondent is responding to.

- **The items lead the respondent**. The items are trying to "educate" or push the respondent toward something. The survey has an agenda.
- **The items and/or response categories are limited in scope**, and thus they exclude some respondents. A good example is the "Neutral/No Opinion" category.
- **The survey is too long**. By the end of the instrument, respondents will tire and stop responding to items.
- **The survey uses so many open-ended items that it is collecting primarily qualitative data and is essentially an interview project**. It will yield a wealth of comments, many of which say very similar things, and may be laborious to read and code.

Please consider using these as litmus tests for the quality of your own future survey work.

How to design a high quality survey

Use these steps, selected from the literature and my own professional experience doing dozens of surveys over the years, to design a high quality survey:

1. Clarify the purpose of your survey.

Begin by considering why choose a survey instead of another method to answer your question. Why is a survey appropriate for your question?

What is the time frame for your survey? Will it be a timely, issue-specific "fact-finding" survey that reveals "How many people think X?" about a specific issue (such as a curriculum adoption, or a bond election)? Will the survey lose its relevance after the moment has passed? Or does your survey aim to measure something ongoing in the culture (like a school climate survey) and thus be used multiple times to build trend data?

Will the data be used to quantify the magnitude of sentiment, attitude, opinion, or behavior? Will the data be used to describe a population? Will the data be used to compare groups on a sentiment, attitude, opinion, or behavior? Or could your data be used to explain which variables are stronger predictors of an outcome than others?

2. Draft a map of the survey.

Designing a good survey is much like designing a good student achievement test. The starting point for a student achievement test is not test questions, it is a map of the different learning objectives. The same goes with a survey. A survey project should begin with a high level list of the overall questions one wants answered.

3. Sample carefully.

What is the sampling method? Is it a convenience sample of people available? If so, what are some sources of sampling bias? What relevant respondents might be left out? What profit might you gain from select a probability sample?

4. Use validated items from other established survey instruments, or write your own high quality items.

Learn from the experts, when possible:

- Writing Survey Questions (The Pew Research Center)
- Best Practices (Washington State University)

Keep survey items short and simple. Avoid long, wordy items that could confuse the respondent.

Avoid double-barreled items. Keep survey items focused on one dimension at a time. (I saw this in education over and over and over again.)

Don't lead or force data from the respondent. Example: Many times I have heard people intentionally withhold a "Neutral/No Opinion" category in order to "force" the respondent to take a stand on an issue. I don't like that practice. If a respondent truly does not understand or have an opinion about a topic, I would rather know that than force the respondent to yield an artificial (and, in my mind, invalid) response.

Allow response categories that span the range of all possible responses. Response categories on a survey item should be **exhaustive** and **mutually exclusive**. This assumes you know the full range of possible responses. If you don't, consider asking this item first as an open-ended item on a pilot survey. Then you can ask it as a closed item on your operational survey.

Be judicious in your use of open-ended items. Allowing respondents to respond in their own words will create a large volume of comments that will take time to read, and many of the comments say similar things. Use open-ended items on a pilot instrument when you don't fully understand an issue and want to see the full range of possible types of responses to it. These types of responses can then become response categories on a closed item on an operational version of the survey.

5. Pilot the questionnaire before going live.

Show the questionnaire to a small sample of intended respondents. Ask them to take the survey, noting the following:

Confusion. Is the purpose of the survey clear to the respondent? Is any part of it confusing to the respondent in any way? Are any items confusing as worded?

Bias. Does the survey truly capture the full scope of respondent experience on the issue? Are some options left out? Do some items lead or force the respondent?

Length. Is the survey an appropriate length? Does the survey tire out respondent? Aim for no longer than 15 minutes.

Validity. Does the survey capture the thinking, (mis)conceptions, ideas, beliefs, sentiments, attitudes, opinions, and/or behaviors you designed it to capture? Or does it also capture extraneous information? Use a "think aloud" method of asking the respondent to verbalize their responses as they take the survey.

In the field, there is not always time or interest to pilot a survey. But in my experience, when possible, piloting has **always** improved the quality of my surveys.

Quasi-Experimental Design

Required reading:

- Chapter 4 (pp. 56-61) in Burkholder et al. (2020)
- Chapter 17 in Burkholder et al. (2020)

To evaluate a program

In education, we often design programs to improve instruction or aspects of schooling to improve outcomes for students. Most programs cost time, money, or other limited resources.

Consider, for the moment, one program that perhaps you have led, implemented, inherited and maintained, or otherwise invested your attention into yourself. Inevitably the question will arise: How well is this program "working"? Do the benefits it yields outweigh the costs? These questions depend on a fundamental question which is our focus for this module:

What counts as evidence? How can we know?

When I worked in districts as the assessment director, questions of program evaluation came up many times. In many cases the originating question was, "Let's look at the data!" and in most cases that really meant, "Some students were part of a program. Let's look at *their* data."

Often the data were scores on a common assessment of student achievement. This included districtwide assessments like DIBELS, STAR, or MAP, or the annual state assessment such as the WASL, the MSP, or the SBA. Students were often selected for a program on the basis of low pretest scores ("Level 1s" and "Level 2s") and the outcome measure was often the same assessment given in a later testing window.

Favorable outcomes for this group then counted as sufficient evidence of program efficacy, and more often than not, the *de facto* evaluator was the person most invested in the program and bent on seeing it continue.

Seldom did it occur to people (or if it did, nobody said anything) what would likely have otherwise happened to these students without experiencing the program. Were they *better off* experiencing this program than the likely alternative? What about very similar students who could have experienced this program but for whatever reason didn't? What were their outcomes?

This whole discussion rests on philosophical assumptions (or commitments, or investments) that we can more or less systematically **cause** better student outcomes and more or less measure this causation. Put another way, it's common to justify the merit or importance of one's work with **causal inferences** that the work **makes a difference**.

Questions about how, when, on whom we collect, organize, and analyze evidence to make causal inferences are questions of **research design**. Design is the framework for a study.

What counts as convincing evidence?

Let's begin by consider a hypothetical scenario of elementary reading, depicted in Table 1. Fifty third grade students score below grade level (40th percentile) on their spring SBA English Language Arts assessment. All of the students return to the same school in the fall for their fourth grade year. Half are assigned to the Innovative Reading Program while the other 25 receive Tier 1 grade level instruction. In the spring, all 50 students take the Grade 4 SBA ELA assessment.

Table 1

Standardized Reading Achievement by Reading Program Placement

N	Pretest	% low income	Placement	Posttest
24	2398 (40th)	51	Tier 1 Grade Level Instruction	2474 (50th)
26	2401 (40th)	49	Innovative Reading Program	2523 (70th)

When the scores become available shortly, they bring good news. All 50 students meet the Level 3 proficiency standard. The average scores of the 25 students in the Tier 1 grade level classroom score is 2474 (roughly the 50th percentile). They've all caught up to grade level. The results of the students receiving the Innovative Reading Program are even better: their average SBA score is 2523, the 70th percentile for fourth grade. Proponents of the program rejoice. "Not so fast!" cry the program critics.

Threats to validity

Here are their objections:

"Of course their scores increased! Reading was a district and school focus. Everyone was talking about it last year. It was in the air." This threat to validity, an alternative explanation for the outcome of the experimental group apart from the treatment itself, is called **history**.

"Of course their scores increased! Kids grow and mature anyway. Studies have shown gains in scores for kids with no formal schooling at all." This threat to validity, an alternative explanation for the outcome of the experimental group apart from the treatment itself, is called **maturation**.

"Of course their scores increased! Having already taken the third grade SBA test, they knew what to expect of the test. They knew how to take it." This threat to validity, an alternative explanation for the outcome of the experimental group apart from the treatment itself, is called **testing**.

"Of course their scores increased! The fourth grade SBA was easier for fourth graders than the third grade SBA was for third graders." This threat to validity is called **instrumentation**.

"Of course their scores increased! Students selected on the basis of extreme low scores will always score higher on average on the posttest because extreme low scores are extreme because of the combined effects of true achievement and measurement error." This threat to validity is called **statistical regression**.

Each of these is a threat to the **internal validity** of a program's treatment. Internal validity is the "basic minimum without which any experiment is uninterpretable: Did in fact the experimental treatments make a difference in this specific experimental instance?" (Campbell and Stanley (1963), p. 5)

Critics raise a couple of additional objections:

"These results are limited. The students selected for the program were less impacted by poverty. They had more favorable demographics. They were different students!" This treat to validity is called **selection bias**.

"These results are limited. The students selected for the program were able to use their advantages to learn at a faster rate." This treat to validity is called **selection-maturation interaction**.

"These results are limited. The students selected for the program were sensitive to the test. They knew they had scored below standard in the spring so they tried harder the next year." This threat to validity is called **reactive** or **interaction effect of testing**.

"These results are limited. The students selected for the program had been low and received more attention and knew that we're watching them closely." This threat to validity is called **multiple-treatment interference**.

These are threats to the **external validity** of a program's treatment. External validity "asks the question of generalizability: To what populations, settings, treatment variables, and measurement variables can this effect be generalized?" Threats to external validity limit the generalizability of the results to broader populations.

What do you make of these objections, in light of the data and what you know of the design of the treatment? Are some more credible than others?

Experimental design with random assignment

The argument for the program is that it dramatically helps struggling readers, which is to say, struggling readers are *better off* in the program compared similar students in Tier 1 classroom instruction. This is because similar students did not gain as much as students in the program. The argument thus hinges on the similarity of the two groups. Any alternative explanation for the improvement of the experimental group must apply to the comparison group.

What if these students were assigned randomly to the conditions? This would have the effect of rendering pre-existing differences not statistically significant (that is, their differences were no more than we would expect to see by chance) … by design. This would strengthen the program advocates' causal argument that (1) they were all the same students and (2) the program worked better than Tier 1 instruction.

Now it is probably wise to let go of it. Assuming we were so inclined, it is seldom feasible to prospectively randomly assign students to conditions, or to keep treatments so cleanly isolated, in real schools, districts, and dioceses. **Nor will it be possible for you to fully design and carry out a prospective experimental design with random assignment in your current doctoral program**.

Quasi-experimental designs

In lieu of a true experimental design, consider adding quasi-experimental designs to your toolbox. These are frameworks for collecting, organizing, and analyzing (primarily quantitative) data for causal inference – such as for program evaluation – that fall short of pure experimental design with random assignment, and therefore expose the evaluation to criticism. As Cox (in Burkholder et al. (2020), 56) puts it well: "The lack of random assignment in quasi-experimental designs means that the groups may not initially be equal or similar. This presents the challenge of ruling out other alternative explanations that could be responsible for any observed outcome." Quasi-experimental designs use one or more work-arounds to mitigate various inescapable threats to validity. Consider three:

The Nonequivalent Control Group Design

This design comes from Campbell and Stanley (1963), who, at that time, claimed:

one of the most widespread experimental designs in educational research involves an experimental group and a control group both given a pretest and a posttest, but in which the control group and the experimental group do not have pre-experimental sampling equivalence. Rather, the groups constitute naturally assembled collectives such as classrooms, as similar as availability permits but yet not so similar that one can dispense with the pretest. (p. 17)

Cook and Campbell (1979) later saw the design as "perhaps the most frequently used design in social science research and is fortunately often interpretable. It can, therefore, be recommended in situations where nothing better is available" (103-4).

Here is the design, and it is the design of the hypothetical example above:

O	X	O
O		O

The design overcomes several of the critics' objections raised above.

History, maturation, testing, and instrumentation are less credible objections because each of these explanations would apply to both groups, and the students in the experimental classroom still outperformed their peers in the comparison classroom. Notice here the added value of a comparison group!

Statistical regression is a valid criticism any time students are selected for a treatment on the basis of extreme scores, because extremely low or high pretest scores will always, on average, regress to the mean on any retest. This should be less of a problem if both groups were selected on the same set of extreme pretest scores.

Interactions between pretesting, selection, maturation, and the experimental treatment mean the students selected for the treatment were aware of their selection and reacted to it. This could be a valid threat whenever students selected for a program become aware of it, especially by virtue of contact with comparison students.

Time series designs

A time series design is, essentially, "the presence of a periodic measurement process on some group or individual and the introduction of an experimental change into this time series of measurements, the results of which are indicated by a discontinuity in the measurements recorded in the time series" (Campbell and Stanley (1963), 37).

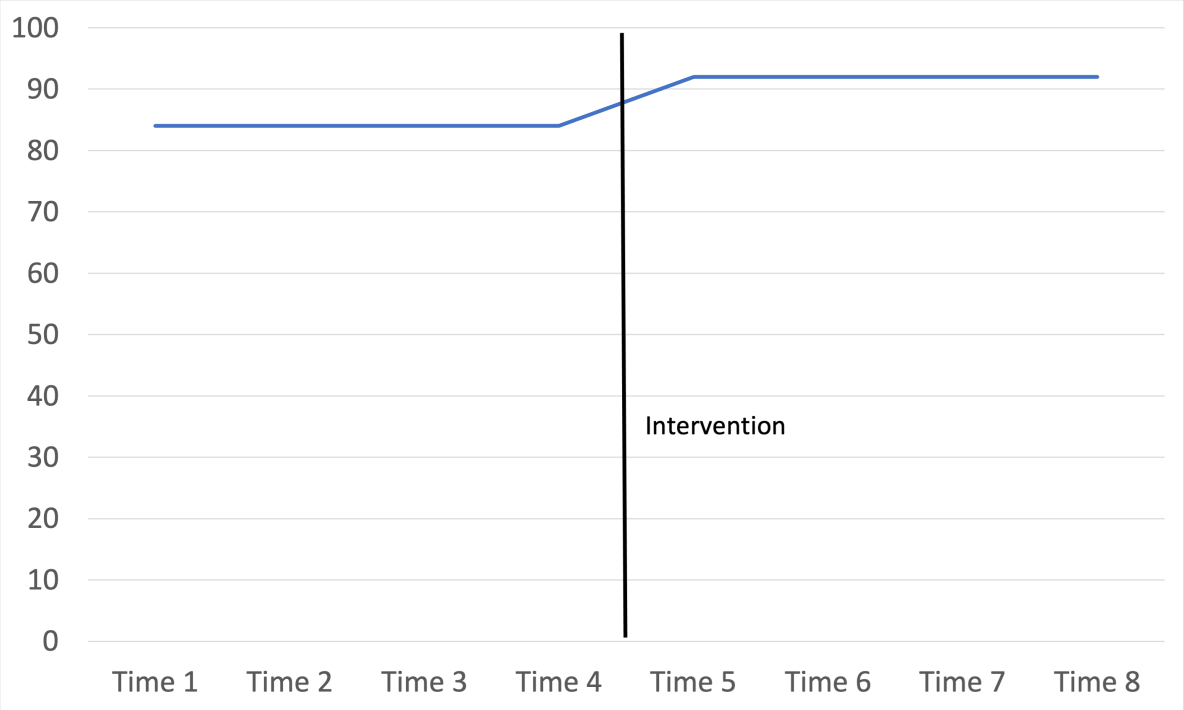
One familiar application of this design might be annual trends in proficiency rates (or average scores) for a grade level at a school.

One group design

A time series design for one group looks like this…

O_1	O_2	O_3	O_4	X	O_5	O_6	O_7	O_8

…and a simple line graph of eight years of results with an intervention between Time 4 and 5 might look like this:



What's going on here? What do we make of these results?

Advocates of the intervention will hail the improvement in scores as evidence of effectiveness.

Are we convinced? What might be some credible threats to validity?

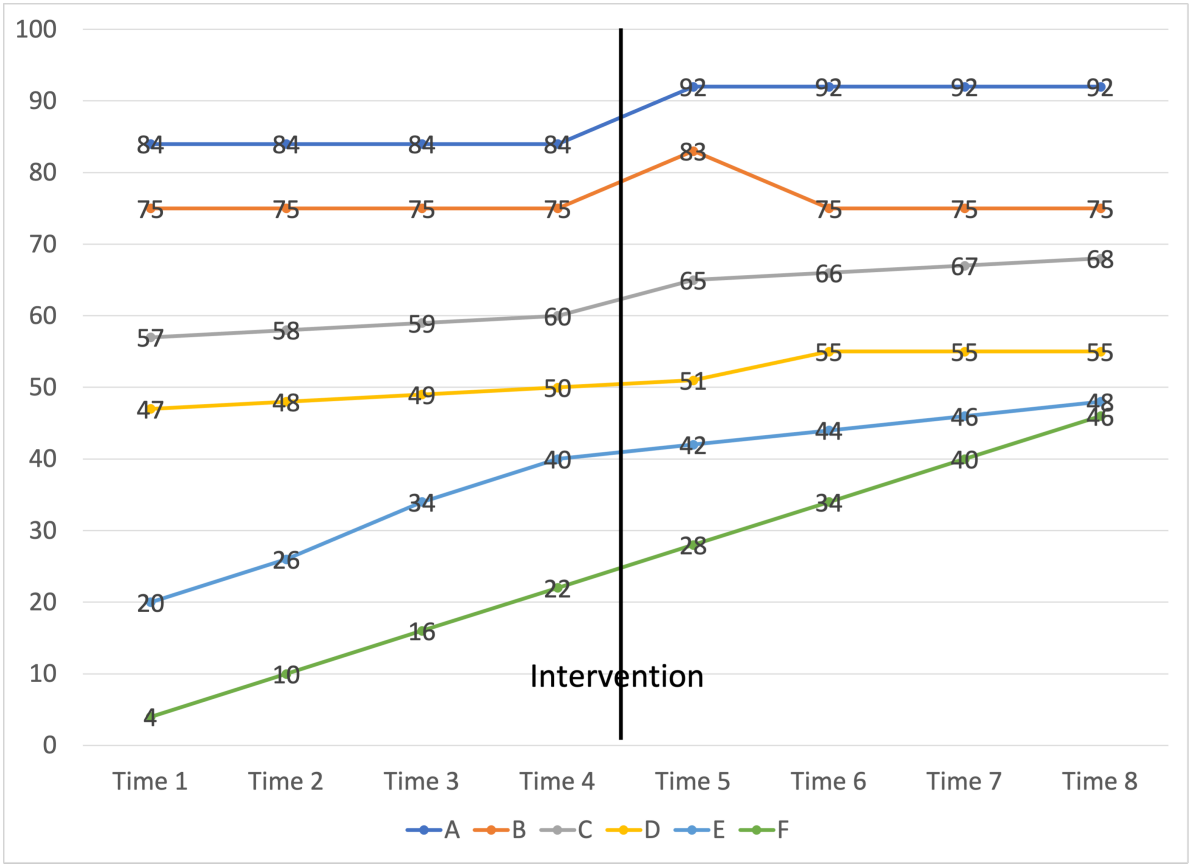
Unless the case is somewhat isolated, this design may be vulnerable to **history** as a rival explanation. An external event could have caused the observed increase.

Multiple-group design

A multiple-group design, as you might imagine, adds groups to the design, like this:

O_1	O_2	O_3	O_4	X	O_5	O_6	O_7	O_8
O_1	O_2	O_3	O_4		O_5	O_6	O_7	O_8

Imagine six groups, all with the same time series of observations. A line graph of their results might look like this:



Improvements since the intervention might look like evidence of the intervention's effectiveness. A larger set of pre-intervention data provides context. In cases that were already improving before the intervention it is more difficult to attribute the improvement to the intervention.

School-level trends in annual aggregate test scores are messy. A school implementing an intervention might be to find schools with comparable demographics to use as controls. But those schools might be doing their own interventions.

Add to this the challenge of sensitivity to instruction.

Seldom do we have so much longitudinal data. More often we have maybe a few years of data, say O_4 and O_5, or Time 4 and Time 5, which is essentially the Nonequivalent Control Groups

Design. In those cases, improvements in scores fall prey to many of the threats to validity outlined above. A time series provides more context and more follow-up.

How to do a program evaluation using a quasi-experimental design

To carry out a retrospective analysis of existing quantitative data using quasi-experimental design for the purpose of program evaluation, you need:

- A group of schools/students/people who experienced some initiative, program, or treatment of interest
- A comparison group of similar schools/students/people who did not experience the program or treatment of interest
- Some pretest data, to establish baseline differences. Maybe these pretest data were the basis for assignment to the initiative or program
- Some demographic data, to explore differences between the groups besides the treatment
- Some posttest data, to establish outcomes of both groups. Ideally the pretest and posttest are the same interest, but this is not required.

Concretely, all this means:

- Excel or other spreadsheet software, and in your worksheets you need:
- A column of pretest scores on a population of schools or students
- Columns for demographic variables on this population of schools or students (gender, race, low income, English language proficiency)
- A column designating which schools or students participated in the initiative or program. Code the schoosl/students receiving the initiative/program as 1, all others 0.
- A column of posttest scores on this population If these data are coming from different places then you need some kind of key variable (such as an ID number) that is common across all the different data sources which you can use to match the data together

Your end game is one spreadsheet where all of these variables are together in one place. Then you can use PivotTable to summarize the data. Keep checking back here for a sample Excel file as a guide.

Recommended further reading

If your current or future work casts you in the role of program evaluator, it may help you to have some additional readings in your library for reference. Here are the three classic texts on experimental, quasi-experimental, and non-experimental design. I personally own and highly recommend all three.

Campbell, D. T., & J. C. Stanley. 1963. Experimental and Quasi-Experimental Designs for Research. Houghton-Mifflin: Boston.

Cook, T.D., & D.T. Campbell. 1979. Quasi-Experimentation: Design and Analysis Issues for Field Settings. Houghton Mifflin: Boston.

Shadish, W. R.., Cook, T. D., & J.S. Campbell. 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin: Boston.

Data Sources

With technological advances in technology, access to data has increased by leaps and bounds, and this includes data available to the public for free download. Here is a curated list of public education data:

Public education data

Washington State public education data (from the Office of the Superintendent of Public Instruction)

- Data & Reporting
- Data Portal
- Washington State Report Card
- Data.WA.gov
- Report Card Spring Assessment Data from 2014-15 to 2021-22

Part II

Qualitative Methods

References

Burkholder, G. J., K. A. Cox, L. M. Crawford, and Hitchcock. 2020. *Research Design and Methods: An Applied Guide for the Scholar-Practitioner*. Sage.

Campbell, D. T., and J. C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Houghton-Mifflin.

Cook, T. D, and D. T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Houghton-Mifflin.