Methods and Instruments for Data Collection

Dr. Jack Huber

2024-08-17

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:

- Pages 56-61 in Cox (2020) (Chapter 4 in Burkholder et al. (2020))
- Milanesi (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 its survival or expansion (for better or worse).

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

Table 2: **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 the following 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).

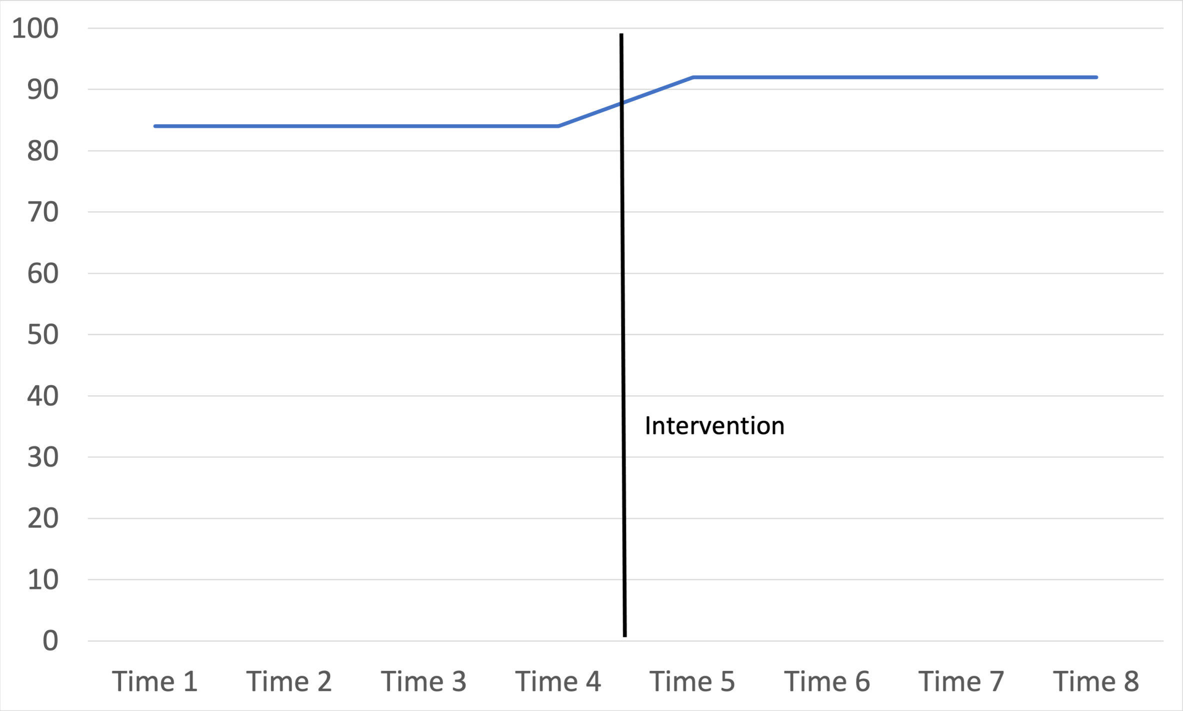
One familiar application of this design might be annual trends in proficiency rates (or average scores) for a grade level at a school.

One group design

A time series design for one group looks like this...

$$\mathrm{O}_1 \quad \mathrm{O}_2 \quad \mathrm{O}_3 \quad \mathrm{O}_4 \quad \mathrm{X} \quad \mathrm{O}_5 \quad \mathrm{O}_6 \quad \mathrm{O}_7 \quad \mathrm{O}_8$$

...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.

Maturation could be a rival explanation if the trend looks like what we would expect of human development, although some kinds of development can probably occur in stages.

Testing is not likely a rival explanation if the same test was used for all observations. It could be a rival explanation if Times 5 through 8 used a different test. This is often an issue in public education when states, districts, and schools adopt different tests from time to time, which gives rise to a saying: “If you want to measure change, don’t change the measure.”

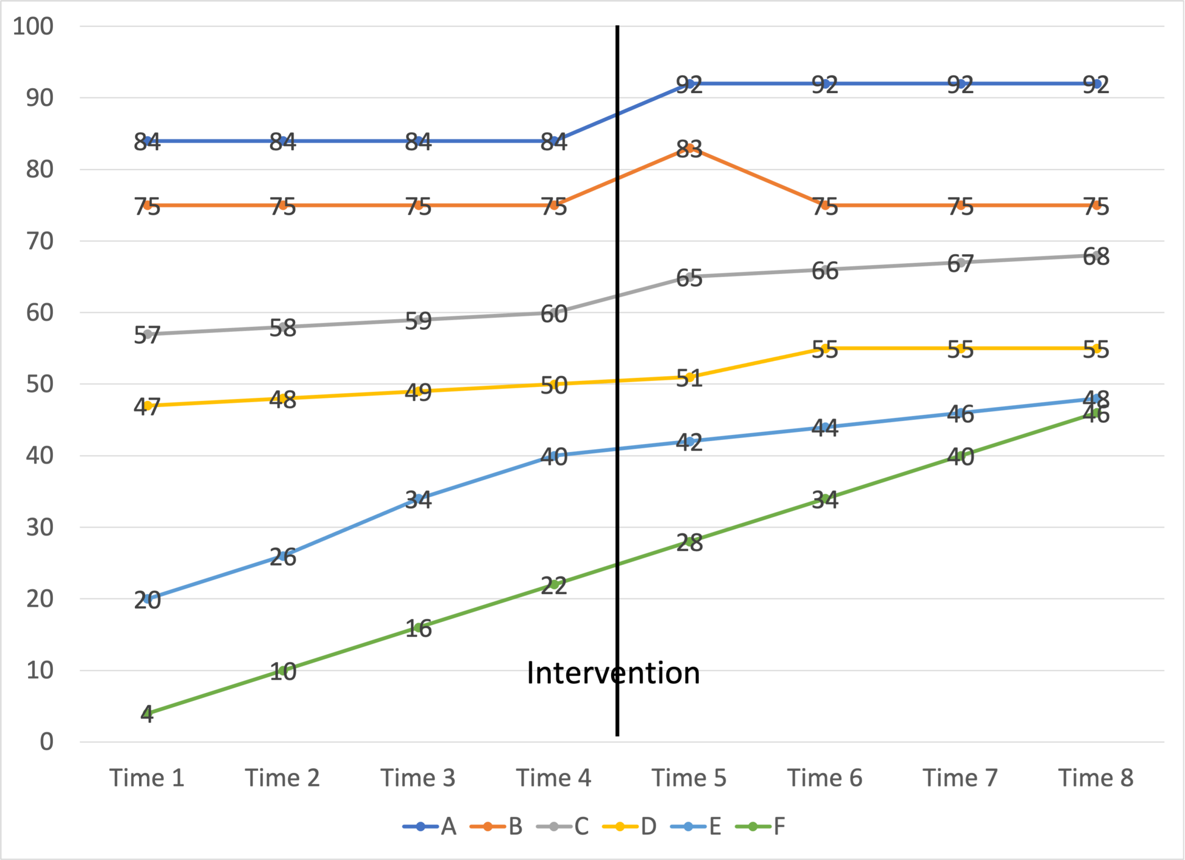
Regression is not a rival explanation because none of the pretest scores began at the extremes.

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 – in this case, schools, – all with the same time series of observations. Imagine an intervention – such as additional discretionary funding, or an additional 1.0 FTE to support struggling learners – that all received or experienced at the same time. A line graph of their results might look like this:



Once again, improvements since the intervention might look like evidence of the intervention's effectiveness. But more data on either side of the intervention provides context. In schools 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 able to find schools with comparable demographics to use as controls. But those schools might be doing their own interventions.

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.

Moral of the Story

Admittedly I've expressed the aforementioned discussion of designs in very abstract, technical terms drawn straight from the methodological literature on this topic (Campbell and Stanley

(1963); Cook and Campbell (1979); Shadish, Cook, and Campbell (2002)).

My intent is far less for you to adopt this language than to understand how this comparative causal logic can apply to programs in your context and/or that have captured your interest and attention.

Compare the Nonequivalent Control Group Design with the Single- and Multiple-Group Time Series Designs. Can you see how much is gained with the addition of groups and measurements on either side of the intervention?

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, or students, or people who experienced some initiative, program, intervention, or treatment of interest at a measurable point in time
- A comparison group of similar schools, students, or 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.

Validity

Required reading:

- Pp. 176-189 in M. S. Stewart and Hitchcock (2020) (Chapter 12 in Burkholder et al. (2020))

Validity is a big deal, in the context of research, action research, or improvement science. Let me put it to you this way: If you're using any kind of instrument to collect any kind of data to make any claim, validity is at stake.

Validity is

- "The quality of being logically or factually sound" (Oxford English Dictionary (Google))
- "the extent to which … inferences and uses of" (Messick (1989))
- "the approximate truth of an inference" … "a judgment about the extent to which relevant evidence support the inference as being true or correct" (Shadish, Cook, and Campbell (2002), p. 34)
- "the best available approximation to the truth or falsity of propositions, including propositions about cause" (Cook and Campbell (1979), p. 37)
- "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (Association, Association, and Measurement in Education (2014))

Many of these definitions are derived from the methodological literature on experimental and quasi-experimental design, and measurement.

Validity is not all or nothing.

In the past, it was common to describe validity as a property of an instrument: "That's not a valid survey." "That's a valid test." Many of my colleagues throughout my career have talked that way, and many times people have wanted me to declare a survey or a test “valid” or “invalid.” They can be forgiven: it’s simple that way. But you won’t hear me make such statements.

Now we understand validity as description of the interpretations and uses of a test, survey, or other instrument; a “judgment of the extent to which evidence and theory support the interpretations and uses of tests” or other instruments of data collection.

We also talk in terms of validation: a process of gathering validity evidence regarding a test. In this sense, when I say validity is not all or nothing, it is instead a judgment based on accumulation of validity evidence.

Types of validity evidence

Measurement experts (those who use instruments like test or surveys to measure abstract concepts like reading comprehension or psychological well-being, respectively) often distinguish several different types of validity evidence.

Content validity

examines the content of the instrument, such as the content of the survey items or test items, in relation to the content of domain it is designed to measure. This kind of question is often raised by students early in life: “Is this on the test?” One way to gather evidence of content validity is to do alignment studies that examine the content of test items in relation to a map of content taught in the classroom.

Construct validity

By construct I mean “the concept or characteristic that a test is designed to measure” (Association, Association, and Measurement in Education (2014), p. 11). Examples include mathematics achievement, general cognitive ability, racial identity attitudes, depression, and self-esteem. Construct validation is a process of scientific study of score meaning. Often this involves statistical studies of data from observer ratings (as in the case of a classroom observation tool) or item responses (as in the case of surveys or tests), and more specifically studies of correlations among ratings, items, total scores, and other measures.

Consequential validity

has to do with the consequences of measurement, particularly in regard to educational assessments designed to support individual growth in skill or proficiency or improvement in teaching. What is the purpose of the assessment? What are its intended consequences? What evidence exists that appropriate use of the assessment actually causes the intended consequences?

An example: A classroom observation tool of preschool teachers. The tool guides observations of preschool teachers and paraeducators at work with students in the classroom. The tool guides what to look for and how to record the observations. The tool is so detailed that multiple individuals using the tool to observe the same classroom at the same time would record the same observations the same way, yielding almost identical data. (This is called inter-rater reliability.) Does use and interpretation of the results produce the intended improvements in instruction? Is the tool diagnostic enough to point out specific areas for improvement?

Validity in regard to causal arguments

Experts on experiment and quasi-experiment distinguish between two different kinds of validity:

Internal validity

"begs the question, 'How truthful is the proposition that a change in one variable, rather than changes in other variables, causes a change in the outcome'?" (M. S. Stewart and Hitchcock (2020), p. 183).

As I described elsewhere, internal validity has to do with isolating the treatment or intervention) as the only or primary explanation of an outcome amid various rival explanations of the outcome.

If you want to make a causal argument that a program produced an intended favorable outcome, you have a burden to prove that the program itself was the primary cause of the outcome. Internal validity is at stake.

External validity

is "the extent to which findings hold true across contexts" (M. S. Stewart and Hitchcock (2020), p. 186).

It could be that a program or intervention did a fantastic job of carefully controlling conditions in order to rule out competing factors. The downside is that is not very realistic, and the favorable results may not apply beyond that setting.

One good way to get a gut sense of external validity is to read a journal article that publishes the results of an experimental study, and by this I mean a true experiment with random assignment. A good example is research on student motivation. More than a few of these studies were classroom experiments where the researcher randomly assigned students to carefully controlled experimental conditions (including a control group), then compared outcomes. These are outstanding rigorous studies that go to great pains to isolate the effect of the treatment on the outcome. One or two is well worth a read to see what all is involved. The unfortunate downside is it is hard to imagine replicating these studies a real classroom. Most ordinary classrooms are too messy (there's too much going on) to achieve such careful control of conditions.

But of all this is admittedly abstract, academic definitions and conceptualizations of validity. It is very "balcony" level.

Let me apply it to a real-world time and place.

Validity issues in context

As I've described elsewhere, I started out as a data analyst in the curriculum and instruction department of a large suburban school district while, on the side, I earned my doctorate (in educational psychology). Then I was an assessment director for ten years. In that span of time, I did a wide variety of analyses of data from surveys, tests, course grades, attendance, discipline, enrollment, and more. I was also responsible for testing students for selection into gifted programs. Here is a series of vignettes of validity issues.

Elementary grades to declare proficiency

At one time, districts spent a lot of money on non-student days for elementary teachers to render numeric grades for students. The purpose of the grades was to convey student proficiency in relation to grade content standards. The scale of the grades ranged from 1 to 4, with 1 and 2 indicating below-standard proficiency, 3 indicating "meets standard", and 4 indicating "above standard" proficiency. These grades populated a standardized standards-based report card that was sent to parents several times per year.

The central validity issue is the extent to which these data supported valid claims about student proficiency. These grades were based primarily on evidence collected in the classroom. With no common grade level rubrics or protocols guided teachers on what evidence to include or how to summarize the evidence that went into the grade, the grades reflected teachers' judgment of student proficiency. Some teachers were more rigorous than others. Some teachers were more attuned to state grade level proficiency standards than others. This means that, in the aggregate, some of the variance in grades was due to variance in teachers (as raters) as well as variance in true proficiency and error variance.

Because students at Grades 3 through 6 were required to take the state assessment in the spring, the state assessment served as an external source of proficiency evidence with which to validate the grades. By matching, for each student, report card grades with state test scores in the same content area, I was able to conduct correlational analysis. To expect perfect correlations would be unrealistic, but strong correlations to the rigorously-validated state assessment would be validity evidence for the report card grades. Weak correlations, however, would pose some challenge to the validity claims based on report card grades.

For a more thorough account of that validity study, see here or here.

Using test scores to select students for gifted programs

Some children are "gifted." They have exceptional cognitive abilities that not well served in the general education classroom. They need to be with other students like themselves in a classroom that can go deeper into topics and cover content at a faster pace than what happens in the general education classroom. Arguably, to deny "gifted" students appropriate curriculum and instruction in a setting such as a self-contained classroom is educational malpractice. Thus every district needs some fair, systematic way to identify which students are gifted in order to place them into the appropriate classroom.

It should be obvious that identification for gifted programs is shot through with validity issues. What exactly do we mean by "gifted", both conceptually and operationally? What counts as convincing evidence? How can districts collect such evidence in fair, consistent ways that will accurately identify which students are "gifted"? Assuming we have valid instrumentation in place, do the identified students actually fare better in the self-contained gifted classroom? Does giftedness depend on socioeconomics? Does it flourish in families with means?

With some guidance from state law, most districts implement some form of identification process that considers different sources of evidence. One source is a standardized test of cognitive abilities called the CogAT. Districts can, with considerable planning and effort, administer this or a similar test to a large population of students and select only the highest scoring as candidates. Critics of standardized testing appropriately point out that testing will misidentify some students, particularly those at the high end. Some especially high achieving student might get lucky and score high, while some gifted students might have a poor testing day and miss a few items, rendering lower than true scores. Parents bent on their children being identified as "gifted" are vigilant about the testing process, quick to challenge any deviation from standard testing procedure that could affect test scores. District gifted personnel help mitigate this problem by triangulating test scores with other evidence, such as that provided by classroom teachers.

Evaluating a state-funded reading assistance program

In what was called the Learning Assistance Program (LAP), the State of Washington provided funding to districts to hire staff members to provide extra instructional support to students who struggling to read at grade level proficiency. Years ago, in a season of scarcity, this program came under scrutiny. “How well is the program working?” “Is it worth the cost?”

For my district, I was able to fashion state assessment results into a multiple-group time series design in order to examine the effect of the program on student reading achievement over time. The results of that analysis, reported here, were that students served by the LAP gained reading proficiency at higher rates than similar students not served by the program over the same period of time. The best available evidence was that LAP was working.

Using test scores to assess readiness for college and career

Readiness for the rigor of college and career has been an increasingly important indicator of the effectiveness of high schools in recent years. What exactly does that mean? If the question “What do these scores mean?” occurs to you – it’s a validity question.

For many years, high school students took aptitude tests like the SAT and ACT to achieve the highest possible scores. The premise was these assessments predicted academic success in the freshman year as evidenced by a strong correlation between SAT and ACT scores and freshman course grades. Correlational work like this is validity research.

You may not know that there is more to this story. That correlational work examining college placement test scores and freshman year course grades also found that the high school grade point average was a stronger predictor of freshman year course grades than the placement test scores – and these models explained only 25% of the variance in freshman year course grades. Twenty-five percent is a lot of variance by social scientific standards, but clearly the vast majority of variance in freshman year course grades had nothing to do with high school grades and test scores.

I wrote an article on this, available to you here (go to Page 10).

Summary

The point of all this is to flesh out what validity means in real world context.

While I have always had a tendency to reify the concept of validity, it is probably more accurate to think of it as a word for describing two things:

- the quality, trustworthiness, defensibility, and/or credibility of data for supporting whatever inference, claim, decision, or use for the data were collected; and/or
- the craftsmanship and care that went into the design of the instrument(s), study, project, or collection of data.

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

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

National

- National Center for Educational Statistics
- U.S. Census Bureau - Education

Other public data

- The General Social Survey (GSS)
- International Consortium for Political and Social Research (ICPSR)
- Pew Research Center
- Data is Plural
- Tableau Public Gallery

Part II

Qualitative Methods

Interviews

Required reading:

- Crawford and Lynn (2020) (Chapter 10 in Burkholder et al. (2020))

Why interview?

Maybe you know, at a high level, that the most interesting questions in your project seem to lead you more to qualitative than to quantitative data. Great! Why might you decide to do *interviews* instead of another qualitative research method?

With interviews you can talk to one person at a time. You can get that person's perspective – in all its richness – in their own words. Maybe that person is the perfect informant; they occupy a position or have had an experience that is critical to your project. Is this person's perspective enough? Maybe that one person's perspective is all you need. What if you had several informants in the same place, talking about the same thing, as in a focus group. Should you do that? Would that give you better data?

This might also be an imminently feasible research method. With interview protocol in hand, you need only reach out to a few individuals and schedule times to meet.

But take care!

Depending on your subject matter, to the extent you want more credible data, this method is a challenge to do well, especially if you have little experience with it. Are your interviewees telling you the full, unvarnished truth? Or are they pulling your chain?

When I worked in districts, I had a saying. Whenever two people went into an office and closed the door, I said, "Well *now* somebody's going to tell the truth!"

If you are in a position of authority, and you are studying your own backyard . . . which is to say you interview people you know, especially people who occupy positions of less or no authority, or whose work you oversee or have influence over, I would, or will, question the veracity of your data.

This does not mean the effort is hopeless. I am willing to be convinced that your data are accurate. I am also willing to accept that you collected the best data you could under real-world constraints. But it does mean that studying your own backyard, which means interviewing people you know, which might include people who report to you, or whose work depends on decisions you make, lays upon you an extra burden of proof and rigor to overcome bias that you would not have, or would have in lesser degree or severity, if you studied a different context and/or people you don't know.

Types of interviews

Crawford and Lynn (2020) distinguish three forms of interviews: structured, semi-structured, and unstructured.

Structured interviews are standardized. They ask the same questions in the same order of all interviewees, as in a phone survey. This offers the obvious advantage that responses are directly comparable across questions. This method might not sound very interesting but it might be appropriate for your project if the kind of data you need is factual and objective, and/or if you do need the same information of every interviewee.

Semi-structured interviews ask the same questions in the same order of all interviewees but have the option to probe for clarity or depth on questions.

Unstructured interviews essentially adapt the data collection to the interviewee. You get your questions in when and where you can, but more important is getting all that this interviewee has to say. This is what I did in my own pre-dissertation qualitative research project. I interviewed high school teachers about their use of data. For that study, I started out with a semi-structured interview protocol, but when I actually got in front of teachers and got them talking, I relaxed my protocol to hear whatever they had to tell me. My thinking was, "What am I missing? Or, how am I missing the boat on this topic?" It was a good experience, and I think they told me the truth. But it presented a challenge in the coding stage when I lacked comparable data from the same questions across all of my teachers.

For that reason and those Crawford and Lynn (2020) state in the chapter, I agree with their advice try using a semi-structured interview.

With one exception: According to Merriam (1998) (p. 75), unstructured interviews are particularly useful

> when the researcher does not know enough about a phenomenon to ask relevant questions. Thus there is no predetermined set of questions, and the interview is essentially exploratory. One of the goals of the unstructured interview is, in fact, learning enough about a situation to formulate questions for subsequent interviews.

Bias

There is a human factor to interviewing.

Maybe that is the appeal. You want to *talk* to people!

Or maybe you don't want to work with numbers or codes in a spreadsheet, or *quantify* anything. You want to work with imagery, perception, intention, reasoning, story, myth, metaphor, identity, emotion, and the like, all conveyed in *words*.

But with the human factor comes the issue of **bias**.

Perceptual biases

Part of the bias issue is how you as the researcher perceive. Crawford and Lynn (2020) rightly point out your perceptual biases as a human being can influence what and how you *take in* information such as:

- what you see in interviews
- what you hear in interviews
- what you find important in interview utterances

To this list I would add: what counts as evidence. If you "are passionate about" or have a deep investment in a program, how open are you to conflicting information?

Researcher biases

There may be aspects of you as the researcher that may hinder your efforts to get the full, unvarnished truth from your informants.

This is the issue that most concerns me with *educator researchers*, especially those **in positions of authority**, or those **studying work in which they are deeply invested**, and those **studying their own backyard**.

Consider these factors:

- Your race, when interviewing someone of a different race, especially about a topic with salient racial inequity
- Your gender, when interviewing someone of a different gender, especially about a topic that involves gender inequalities

- Your position of authority, when interviewing someone in a position of less authority; or someone whose work is affected by decisions you made; or worse, **someone who reports to you**
- Your energy, perhaps as an extrovert, when interviewing an introvert
- Your Type A personality, when interviewing someone who is Type B
- Your passion for a topic, when interviewing someone who does not share your passion for it

I have no doubt that there are different perspectives on bias from more sophisticated qualitative researchers and I encourage you to seek out such information.

In past student work I have appreciated acknowledgment of bias: “Yes, I have a bias in this study. I have a history with this topic and am passionate about it.”

But, to me, bias is not a flag to fly proudly, but a potential threat to the credibility of the data that a conscientious interviewer should try to mitigate. If you want to interview me, and I can tell from your position, your voice, your face, your body language, your questions, your tone of voice, that you “are passionate about this topic” **I will be very careful what I say to you, and how I say it**. As a sensitive person, I may not want to say anything to hurt your feelings. I may be more inclined to tell you what I think you want to hear. I may withhold information or emotion that may be difficult for you to hear.

Mitigating bias

What steps can you take to acknowledge and work to mitigate the effects of these factors on the credibility of your data?

If nothing else, the first thing I recommend is to acknowledge any and all biases that could affect both your own perceptions as well as your interviewees’ perceptions. Even to make them conscious like this could help you mitigate them before, during, and after data collection.

As a starting point, please carefully study Table 10.1 in Crawford and Lynn (2020) which offers a helpful array of different biases and ways to mitigate them. Among these I emphasize:

- “Find a comfortable, neutral facial expression and maintain it. if you tend to be naturally facially expressive, practice managing that in practice interviews. You want to be very careful not to express surprise, agreement, pleasure, or offense in reaction to the interviewee”
- “Limit nodding”
- “Audiotape or document verbatim. Listen to transcripts following the interview and/or review the notes”

- "It is best not to interview people you are connected to in some way"
- "Audio record interviews, and be prepared to document verbatim on site if the interviewee refuses to be audiotaped"
- "Start with more innocuous questions–such as demographic, factual-type questions–and build to the deeper questions"

To all these in Table 10.1 I would add:

- Be absolutely clear in your consent process that participation at any level is voluntary and can be withdrawn at any time and data will be anonymous
- Seriously consider interviewing people you do not know
- Seriously consider collecting data in a different organization
- Be prepared to **bracket** ("a method used in qualitative research to mitigate the potentially deleterious effects of preconceptions that may taint the research process (Tufford and Newman (2010), p. 80). There are several ways to bracket, including memoing, reflexive journaling, and external interviews (Tufford and Newman (2010))")

Tips for conducting good interviews

Crawford and Lynn (2020) offers sage advice on the kinds of details to consider when conducting an interview. I reproduce and comment on the following:

Be clear on the difference between your research questions and your interview questions.

They are not the same!

Maybe your research questions are very simple and straightforward. More likely, your research questions are loaded with more conceptual and/or theoretical background from your reading and literature review than your interviewees know. To ask your interviewees your research questions straight up would be too big for them.

My advice is to not ask the research question directly. Instead, ask a lot of smaller background questions that will amount to an answer to the big research question.

Do a practice interview

> "Practice interviewing with a friend or colleague, and record the practice sessions on videotape. Obtain feedback and conduct a self-critique with regard to your posture, your manner of asking questions, and any subtle facial expressions or voice tone that might influence participant responses. The goal here is to learn if, for example, you are asking questions too quickly or using body language to communicate judgment. The goal can also be to learn if you are missing opportunities to probe more deeply into responses" (152).

This is excellent advice. I strongly encourage you to practice an interview seeking this kind of feedback before doing more interviews.

Take steps to acknowledge and mitigate bias

> "When conducting the interview, the researcher must avoid body posture, body language, voice tone, and linguistic constructions that communicate judgment or lead the participant. For example, a physical response that communicates 'I like what you said' would inject researcher bias into the interview. Likewise, asking a follow-up probe such as 'Don't you think people would be better off if . . . ?' communicates an opinion on the part of the researcher, whereas a more open question such as 'What do you think would be better . . . ?' allows the respondent to more freely express opinion" (153).

This, too, is excellent advice.

Try using these four kinds of questions . . .

Merriam (1998) presents four kinds of questions that work well in interviews:

Table 6: Four Types of Questions that Work Well in Interviews

Type of Question	Example
Hypothetical Question: asks what the respondent might do or what it might be like in a particular situation; usually begins with "What if" or "Suppose"	"Suppose it is my first day in this training program. What would it be like?"
Devil's Advocate Question: challenges the respondent to consider an opposing view	"Some people would say that employees who lose their job did something to bring it about. What would you say to them?"

Type of Question	Example
Ideal Position Question: asks the respondent to describe an ideal situation	“What do you think the ideal training program would be like?”
Interpretive Question: advances tentative interpretation of what the respondent has been saying and asks for a reaction”	“Would you say that returning to school as an adult is different from you expected?”

. . . and avoid using these three kinds of questions

By contrast, Merriam (1998) illustrates three kinds of questions you should avoid in interviews:

Table 7: Questions to Avoid

Type of Question	Example
Multiple Questions	How do you feel about the instructors and the classes?
Leading Questions	What emotional problems have you had since losing your job?
Yes-or-No Questions	Do you like the program? Has returning to school been difficult?

Sample Interview Protocols

Here are some sample interview protocols:

- An interview study of small schools in Washington State
- My interview study of high school teachers using assessment data in Washington State

Case Selection

Required reading:

- Pages 88-90 in Crawford (2020)

"Qualitative studies … are not concerned with representing a population but, instead, are focused on relevance to the research question" (Crawford (2020), p. 88).

Whoa!

Did you catch that?

When consuming research, how often do we concern ourselves with matters of sample size and representativeness . . . on the premise that larger samples are better?

In qualitative research, the concern is not for generalizing to a population. As I've suggested elsewhere in this course, whereas quantitative research generally tries to test what is generally true across a population, qualitative research generally tries to discover what is deeply true of a small number of cases.

Great!

Then how do we go about selecting cases for a qualitative study? What cases do we select? And how many?

Crawford (2020) echoes other qualitative researchers (Merriam (1998)) in saying that qualitative case selection is **purposive** or **purposeful**. It is based on *relevance* rather than representativeness.

So how do you go about purposive sampling? Here is some guidance from different sources:

Establish criteria

This means you begin with your all-important research question. It should be specific enough to offer clues about what counts as relevant data, and where to find them. From there you should be able to spell out some eligibility criteria.

Are you studying parent engagement? Great! What parents qualify? Any parents? Engaged parents? Disengaged parents? Do you need both to study the contrast? And what counts as engagement?

Are you studying high school student engagement in learning? Great! What does that mean in concrete terms? Which students? Engaged students, or disengaged students? Or do you need both to study the contrast? And what counts as engagement? If you walk into a classroom and see a student bored out of their mind, what if that person is getting excellent grades? Does disengagement necessarily mean low academic achievement?

Crawford (2020) rightly points out that you need to define your terms. You need specificity in your research question. Problem of practice needs to become . . .

- **this specific problem**,
- **of *this kind* of practice**,
- **here *on the ground*, in *this specific time and place***,
- with ***these people***.

Once you have identified and spelled out eligibility criteria, here are several methods of carrying out purposive sampling:

- Snowballing - asking the current participant for a referral to a next participant
- Convenience - using those who are readily available
- Opportunistic - capitalizing on unexpected leads

All three of these methods might be problematic for sampling for quantitative data which needs to be faithful to probability theory. But these methods make more sense from the standpoint of fidelity to theory, and to the purpose of getting the real story.

Go for saturation

How many cases or samples of data to collect? Crawford (2020) points to the concept of **saturation**, by which they mean:

> (1) Continued analysis yields no new information and
>
> (2) there are no unexplained phenomena (Crawford (2020), p. 90)

Do I need more than one?

I'll add one more consideration. Assuming you have a well-defined research question derived from your lit review and conceptual work and are ready to sample cases, go minimal.

Maybe you've identified the perfect case of your study. Maybe it's a program that exemplifies your project. Or an innovative high school attempting to do all the right things. Or a school that has figured out how to do PLCs effectively. Or a "beating the odds" school with high poverty and high achievement. Or a key person within reach who is directly implicated in your problem of practice and research question.

Do you need more than one?

How many more?

Two?

This might be a very effective way to think about sampling for your qualitative project.

Focus Groups

Required reading:

- Pp. 148-150 in Crawford and Lynn (2020)

Why do a focus group?

A focus group is a group of individuals – Crawford and Lynn (2020) recommends 6-10, D. W. Stewart and Shamdasani (1990) suggests 8-12 – convened to discuss your research topic. The term “focus” means discussion is limited to a small number of issues.

As with interviewing, you don’t actually ask them your research question straight up. You ask them smaller questions more anchored in their experience that cumulatively lead up to your research question or provide a body of data that address your research question.

Burkholder et al. (2020) has little to say about focus groups. Crawford and Lynn (2020) gives the method brief discussion in her chapter on interviewing. She offers several criteria for deciding whether to do a focus group:

- **Study design**. Qualitative research traditions differ epistemologically in what counts as data and in how data should be collected. Data that emerge from a focus group have the character of a group interaction, or conversation, in which group dynamics are at work. Such might make little sense if your the tradition grounding your research emphasizes individual narrative or individual experiences that might be sensitive. On the other hand, what if your research question is about some aspect of shared culture? – of shared practice, belief, organizational ritual, collective identity, theory of action? In that case a focus group might be a good way to go.
- **Practical considerations**. It might be easier to arrange a focus group than a series of interviews with individuals. And the data you get from a focus group could be a one and done. Considerations of accessibility, privacy, freedom from distractions, and ease of high quality data collection also play a role. In this short time frame of your program, you would be wise to consider the precious time you have to collect and analyze data.

- **Group talk**. In a focus group, more vocal participants can stimulate thinking and discussion in the group that might not otherwise occur in individual interviews. But these more vocal individuals might also dominate discussion and, in the absence of skillful moderation, take the discussion in unhelpful directions. I have witnessed focus groups essentially become gripe sessions.

D. W. Stewart and Shamdasani (1990) suggests that focus groups

- "are particularly useful for exploratory research where relatively little is known about the phenomenon of interest" (15).
- "are also useful following analysis of data from large, quantitative survey. In this latter use the focus group facilitates interpretation of quantitative results and adds depth to the responses obtained in the more structured survey" (15).

The authors go on to enumerate several advantages:

Advantages of Focus Groups

1. "Focus groups provide data from a group of people much more quickly and at less cost than would be the case if each individual were interviewed separately. They also can be assembled on much shorter notice than would required for a more systematic, larger survey.
2. Focus groups allow the research to interact directly with respondents. This provides opportunities for the clarification of responses, for follow-up questions, and for the probing of responses. Respondents can qualify responses or give contingent answers to questions. In addition, it is possible for the research to observe nonverbal responses such as gestures, smiles, frowns, and so forth, which may carry information that supplements (and, on occasion, even contradicts) the verbal response.
3. The open response format of a focus group provides an opportunity to obtain large and rich amounts of data in the respondents' own words. the researcher can obtain deeper levels of meaning, make important connections, and identify subtle nuances in expression and meaning.
4. Focus groups allow respondent ot react to and build upon the responses of other group members. This synergistic effect of the group setting may result in the production of data or ideas that might not have been uncovered in individual interviews.
5. Focus groups are very flexible. They can be used to examine a wide range of topics with a variety of individuals and in a variety of the settings.
6. Focus may be one of the few research tools available for obtaining data form children or from individuals who are not particularly literate.

7. The results of a focus group are easy to understand. Researchers and decision makers can readily understand the verbal responses of most respondents. This is not always the case with more sophisticated survey research that employs complex statisitical analyses" (16).

But focus groups also have their limitations:

Limitations of Focus Groups

1. "The small numbers of respondents that participate even in several different focus groups and the convenience nature of most focus group recruiting practice significantly limit generalization to a larger population. Indeed, persons who are willing to travel to a locale to participate in a one- to two-hour group discussion may be quite different from the population of interest, at least on some dimension, such as compliance or deference.
2. The interaction of respondents with one another and with the researcher has two undesirable effects. First, the responses from members of the group are not independent of one another, which restricts the generalizability of results. Second, the results obtained in a focus group may be biased by a very dominant or opinionated member. More reserved group members may be hesitant to talk.
3. The 'live' and immediate nature of the interaction may lead a researcher or decision maker to place greater faith in the findings than is actually warranted. There is a certain credibility attached to the opinion of a live respondent that is often not present in statistical summaries.
4. The open-ended nature of responses obtained in focus groups often makes summarization and interpretation fo results difficult.
5. The moderator may bias results by knowingly or unknowingly providing cues about what types of responses and answers are desirable" (D. W. Stewart and Shamdasani (1990), 17).

Quality of Qualitative Research

Required reading:

- M. S. Stewart and Hitchcock (2020) (Chapter 12 in Burkholder et al. (2020))
- Pages 277-280 in @Miles-Huberman-1994

Oceans of ink have been spilled on what counts as high-quality quantitative data. Some of the ink has been my own. You might say I have a thing about validity. Here are some concrete examples:

- A survey researcher collects a carefully designed random sample to best represent the target population.
- An experimental researcher goes to pains to design experimental conditions to rule out rival explanations of the effect of a treatment on an outcome measure.
- A measurement expert employs sophisticated statistical techniques to gather evidence that the responses from a standardized instrument validly measure the construct the instrument was built to measure.

In each case, big decisions might be at stake. Leaders paid big salaries to make big decisions about large sums of money and many other peoples' work need solid information.

Maybe, for whatever reason, quantitative research holds no appeal to you and you prefer qualitative work. Great! I've seen enough good qualitative research to have immense respect for the methodology.

But if you suspect qualitative research to be any easier . . . any less rigorous . . . any *more do-able* . . . concerns of validity, reliability, dependability, and trustworthiness any less important . . . I'm sorry to disappoint you. Heavy is your burden to prove your qualitative work is credible. As Merriam (1998) puts it:

> … how can consumers of research know when research results are trustworthy? They are trustworthy to the extent that there has been some accounting for their validity and reliability, and the nature of qualitative research means that this accounting takes different forms than in more positivist, quantitative research. 198

Common Challenges

Qualitative research will likely involve talking to people, as in an interview or focus group. How do you know that someone is telling you the truth? Or are they telling you a *version* of the truth – a sanitized or guarded truth?

When I read your finished product, why should I, as the reader, trust what you're telling me? Merriam (1998) offers a set of critical questions about the trustworthiness of your research that you might prepare to address:

Table 8: Challenging the Trustworthiness of Qualitative Research (from Merriam (1998))

1. What can you possibly tell from an *n* of 1?
2. What is it worth just to get one person's interpretations of someone else's interpretation of what is going on?
3. How can you generalize from a small, nonrandom sample?
4. If the researcher is the primary instrument for data collection and analysis, how can we be sure the research is a valid and reliable instrument?
5. How do you know the researcher isn't biased and just finding out what he or she expects to find?
6. Doesn't the researcher's presence so alter the participant's behavior as to contaminate the data?
7. Don't people often lie to field researchers?
8. If somebody else did this study, would they get the same results?

In qualitative research, the instrument is the researcher, and subjectivity and bias are present at all levels:

- What counts as a problem of practice? For that matter, what counts as a problem? What you consider a problem may not seem so problematic to me. It's a matter of perspective – of bias.
- What research questions are interesting? I might not find them so interesting. Or I might frame different research questions of the same topic. In interest lies bias.
- What questions will you ask of your participants? Why those questions and not others? Still more bias.

- How will you analyze the data? What categories or themes will you code from all your verbal data? Bias again.
- What conclusions will you write up from your data? Are you sure about those conclusions? All roads lead to . . . bias.

As you would in quantitative study, you can ask critical questions of different parts of the qualitative research study: “Were the interviews reliably and validly constructed; was the content of the documents properly analyzed; do the conclusions of the case study rest upon data?” (Guba and Lincoln, 1981, p.375, quoted in Merriam (1998)).

Strategies for Addressing Challenges

These are all important considerations, and I’ll continue to harp on them. But there are strategies you can use to deal with some of these challenges.

Merriam (1998) further offers six strategies to help you address the internal validity (not an ideal term) of your qualitative research:

Table 9: Six Strategies for Addressing Internal Validity of Qualitative Research (from Merriam (1998), p. 204-5)

1. **Triangulation**. Use multiple investigators, multiple sources of data, and/or multiple methods to confirm emerging findings, not as technological solution for ensuring validity, but for “holistic understanding.”
2. **Member checks**. Take data and tentative interpretations back to the people from whom they were derived and ask them if the results are plausible.
3. **Long-term observation**. Gather data over a period of time in order to increase the validity of the findings.
4. **Peer examination**. Ask colleagues to comment on the findings as they emerge.
5. **Participatory or collaborative modes of research**. Involve participants in all phases of the research from conceptualizing the study to writing up the findings.
6. **Researcher’s biases**. Clarify your assumptions, worldview, and theoretical orientation at the outset of the study.

Another somewhat erroneous challenge to qualitative research is its reliability – that is, if someone else were to conduct the same study, would they get same results. That’s a reasonable question in quantitative research. But not in qualitative research. Merriam (1998) suggests that

> rather than demanding that outsiders get the same results, a researcher wishes outsiders to concur that, given the data collected, the results make sense–they are consistent and dependable. The question then is not whether findings will be found again but whether the results are consistent with the data collected. 206>

Here are several techniques qualitative investigators can employ to ensure that results are dependable.

Table 10: Strategies to Enhance Dependability of Qualitative Research (from Merriam (1998), p. 206-7)

1. **The investigator's position**. "The investigator should explain the assumptions and theory behind the study, his or her position vis-a-vis the group being studied, the basis for selecting informations and a description of them, and the social context from which data were collected" (LeCompte and Priessle, 1993; quoted in Merriam (1998), 206-7).
2. **Triangulation**. Again, use multiple methods and data sources to triangulate findings.
3. **Audit Trail**. Subject yourself and your research to audit. Put it all out there. Describe in detail how you conducted your study – how data were collected, how categories were derived, how decisions were made throughout the inquiry – so another can audit your work. (That's right.)

Let's talk about **generalizability**.

My sense is that most educators have at least a simple understanding of this concept: research results should apply to more people or settings than just those few, from a particular time and place, who provided data. And, taking it a step further, perhaps you have seen, heard, or even voiced challenges to the applicability or generalizability of some research.

If you ask qualitative research to generalize to a broader population, or if you uncritically assume your own qualitative study will generalize to a broader population, then you would be wise to reconsider. Qualitative research, usually based on small samples, is not designed, and arguably does not intend, to generalize to a population. As I've said elsewhere: Quantitative research aims to discover or test what is generally true of a population, while qualitative research aims to find what is deeply true of a small population.

Merriam (1998) describes several alternative ways to think about broader applicability of qualitative research findings. One is **concrete universals** (Erickson, 1986). The idea here is that "the general lies in the particular; that is, what we learn in a particular situation we can transfer to similar situations subsequently encountered" (Merriam (1998), 210), much like we do in everyday life. A second is **naturalistic generalization**. "Drawing on tacit knowledge, intuition, and personal experience, people look for patterns that explain their own experience as well as events in the world around them. 'Full and thorough knowledge of the particular' allows on to see similarities 'in new and foreign contexts' (Stake, 1978, 6). A third is **reader**

or user generalizability. This means”leaving the extent to which a study’s findings apply to other situations up to the people in those situations” (211).

Merriam (1998) goes on to suggest several strategies for helping qualitative research be more widely applicable:

Table 11: Strategies for Enhancing Applicability of Qualitative Research (from Merriam (1998), p. 211-12)

1. **Rich, thick description**. Providing enough description so that readers will be able to determine how closely their situations match the research situation, and hence, whether findings can be transferred.
2. **Typicality or modal category**. Describing how typical the program, event, or individual is compared with others in the same class, so that users can make comparisons with their own situations.
3. **Multisite designs**. “Using several sites, cases, situations, especially those that maximize diversity in the phenomenon of interest; this will allow the results to be applied by readers to a greater range of others situations.

Miles and Huberman (1994) also take up this issue of trustworthiness of qualitative research. As you might imagine, in the qualitative research world there are multiple perspectives on what counts as criteria for judging the quality of qualitative research, but Miles and Huberman (1994) do agree that

> Our view of qualitative studies take place in real social world, and can have real consequences in people’s lives; that there is a reasonable view of ‘what happened’ in any particular situation (including what was believed, interpreted, etc.); and that we who render accounts of it can do so well or poorly, and should not consider our work unjudgable. In other words, shared standards are worth striving for (Howe and Eisenhart (1990); Williams (1986)).

They lay out several categories of quality and list for each a set of questions for shedding light on these categories. Read these few pages and let these questions guide your work.

References

Association, American Educational Research, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association.

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.

Cox, Kimberley A. 2020. "Quantitative Research Designs." In *Research Design and Methods: An Applied Guide for the Scholar-Practitioner*, edited by Gary J. Burkholder, Kimberley A. Cox, Linda M. Crawford, and John H. Hitchcock, 51–65. Sage Publications.

Crawford, Linda M. 2020. "Qualitative Research Designs." In *Research Design and Methods: An Applied Guide for the Scholar-Practitioner*, edited by Gary J. Burkholder, Kimberley A. Cox, Linda M. Crawford, and John H. Hitchcock, 81–98. Sage Publications.

Crawford, Linda M., and Laura Knight Lynn. 2020. "Interviewing Essentials for New Researchers." In *Research Design and Methods: An Applied Guide for the Scholar-Practitioner*, edited by Gary J. Burkholder, Kimberley A. Cox, Linda M. Crawford, and John H. Hitchcock, 147–59. Sage Publications.

Howe, K. R., and M. Eisenhart. 1990. "Standards for Qualitative (and Quantitative) Research: A Prolegomenon." *Educational Researcher* 19 (4): 2–9.

Merriam, Sharan B. 1998. *Qualitative Research and Case Study Applications in Education*. Jossey-Bass Publishers.

Messick, Samuel. 1989. "Validity." In *Educational Measurement*, edited by Robert L. Linn, 3rd ed., 13–103. MacMillan Publishing: American Council on Education.

Milanesi, Louis. 2020. "Program Evaluation." In *Research Design and Methods: An Applied Guide for the Scholar-Practitioner*, edited by Gary J. Burkholder, Kimberley A. Cox, Linda M. Crawford, and John H. Hitchcock, 259–74. Sage Publications.

Miles, Matthew B., and A. Michael Huberman. 1994. *Qualitative Data Analysis: An Expanded Sourcebook*. Second. Sage.

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

Stewart, David W., and Prem N. Shamdasani. 1990. *Focus Groups: Theory and Practice*. Vol. 20. Applied Social Research Methods Series. Sage Publications.

Stewart, Molly S., and John H. Hitchcock. 2020. “Quality Considerations.” In *Research Design and Methods: An Applied Guide for the Scholar-Practitioner*, 175–89. Sage Publications.

Tufford, L., and P. Newman. 2010. “Bracketing in Qualitative Research.” *Qualitative Social Work* 17 (1): 80–96.

Williams, D. D. 1986. “Naturalistic Evaluation: Potential Conficts Between Evaluation Standards and Criteria for Conducting Naturalistic Inquiry.” *Educational Evaluation and Policy Analysis* 8 (1): 87–99.