Introduction to Research Design OR "and *how* did they measure that?" OR Correlation doesn't imply causality[1]

Elana Broch, Ph.D., MLIS
Stokes Library
Princeton University
ebroch@princeton.edu

Introduction

- Focus on concepts and vocabulary of research design
- Designed for consumers of research to help you develop a healthy skepticism when you hear research findings
- Focus on quantitative research (i.e., not focus groups or case studies)
- Focus on non-experimental research (no laboratory experiments) although most of what we will cover would apply.

Examples of Types of Research

Experimental

1. *Laboratory experiment*
2. Clinical Trials—As close to laboratory as possible
    a. Used to evaluate medical treatment
    b. Random assignment of participants to treatment groups

Observational

1. Decennial Census, Economic census, American Community Survey
    a. Attempts to collect data from population
    b. Doesn't match people up over time
2. Ongoing (longitudinal) government surveys
    a. NLSY, NELS88, American Community Survey

These types of research studies may or may not involve experimental "manipulation"

1. Evaluation Research ("What Works")
    a. Is a two-hour 'Introduction to Research Design' course an effective way to teach librarians about research design?
2. Opinion Surveys: presidential approval rating, presidential approval rating after a discussion of the war in Iraq
3. Market research: "Pepsi Challenge"

[1] Presented at the annual meeting of the Special Libraries Association, June 15, 2008, Seattle, WA.

Part I: Variables and constructs

Variables

A variable is something that varies.
We are usually interested in looking at more than one variable at a time.

Physical variables

1. Probably easiest type of variable to work with
    a. height, weight, age, gender, BP
2. Usually unambiguous, but e.g., BP and white coat.
3. Requires articulating how the variables will be measured.
    a. Goal is consistency; obtained through specificity. Example: weight in kg before breakfast, without shoes.
4. When self-reported, requires participant to know the answer AND report honestly

Demographic variables

1. For example: occupation, education, marital status, religion.
2. These are constructs and require articulating how the variables will be measured. Articulating=operationally defining.
3. Operational definition will depend on the study. Example: marital status may need to be broadened to accommodate "living together." Do you care about how long the person has been married? If so then the question should be about "years married."

Achievement variables

1. If you are trying to measure what someone has learned assume the respondent answers to the best of their ability.
2. You cannot give the identical question again to measure knowledge unless a long period of time elapses.
3. For example: SAT,GRE,NAEP

Opinions/attitude variables

Who will you vote for in the Presidential election? How many children did you want to have? Is a two-hour 'Introduction to Research Design' course an effective way to teach librarians about research design?

1. Wholly dependent on study participant's interpretation; Very idiosyncratic
2. Requires respondent to be honest
3. Requires respondent to remember accurately if retrospective
4. Suffers from lack of true "scale"

5. Opinions/attitude variables are highly dependent on question design
    a. Four-point vs. five- point scales
        i. Strongly agree, agree, disagree, strongly disagree
        ii. Strongly agree, agree, neutral, disagree, strongly disagree
    b. Yes/No Questions
    c. On a scale from 1 to 10….
    d. Can be “open-ended” (e.g., what do you hope to learn from today’s class?)
        i. Open-ended questions usually require coding before they can be summarized.

Constructs and Operational Definitions

Same construct (concept) can be measured in many ways

1. When we “operationally define” “constructs” we articulate the variables we will use to measure the construct.
    a. For example, “Media exposure” and “aggression”
        i. What media to include?
        ii. Self-report vs. Nielsen meter?
        iii. Number of hours in previous day? Week? Month?
        iv. Does all TV count?
2. Pros and cons of controlled vocabulary parallel those for operational definitions.
    a. PRO: Two catalog users using the same controlled vocabulary should come up with the same list of books, but
    b. CON: Loss of information going from construct to operational definition parallels the loss of information going from the content of a book to its LCSH.

Importance of Operationally Defining Constructs

1. Researchers (and the consumers of their research) often forget that just because they *think they’re measuring “obedience,” “parental involvement,” or “socioeconomic status,”* that we are only measuring a proxy for these constructs.
2. From now on, when you hear someone describing a study, the first thing that should come into mind is “and how did they measure that?”

Levels of measurement

When deciding what variables to use you need to be aware of their **level** of measurement

1. **Categorical variables:** non-numeric
    a. Race, major, country of origin.
    b. Often can form the basis for comparison groups.
    c. Can be represented by a number, but numbering is arbitrary.
2. **Ordinal variables**: non-numeric but order has meaning
    a. survey responses (e.g., 5- point scale);Academic rank; lung cancer stages.
    b. If numbered, numbers reflect the order but are not “real.”
3. **Interval and ratio**: numbers have meaning.
    a. Interval and ratio are usually treated the same for purposes of analysis.
    b. Example of interval (temperature). Arbitrary zero
    c. Examples of ratio (number of years smoked, time spent on homework, hours exposed to media violence).

Level of measurement determines what statistical analyses you can do.

1. Use interval/ratio, if possible. You can convert ratio to ordinal or nominal, but the reverse is not true.
    i. How many hours a week do you spending watching television?
    ii. How frequently do you watch television?
        1. Very often, often, sometimes, never
        2. Very often, often, sometimes, never depends
    iii. Do you watch a lot of television? Yes/No

Summary of Part I: Variables and Constructs

Typologies of Variables

1. Physical, demographic, achievement, opinion/attitudinal
2. Levels of measurement: nominal, ordinal, interval/ratio

Constructs (concepts) and operational definitions”

- “and how did you measure that”

PART II: Typologies of research or "How big is your grant?"

Typologies covered:

1. Goal: Descriptive, Relational, or Causal
2. Degree of Experimental control: Observational, Quasi-Experiment, Experiment
3. Time: Prospective or Retrospective
4. People: Sample or Population
5. Data Collection: Longitudinal or Cross-sectional

All research can be classified in terms of these characteristics or a combination of them. Sometimes decisions are made for sound research-based reasons; time and financial constraints enter into the decision as well.

Research Typology number 1: Goal of Research

1st **Descriptive** (Incidence of lung cancer is increasing) can lead to

2nd **Relational** (People who have lung cancer seem to be smokers) can lead to

3rd **Causal**…Smoking **causes** lung cancer…

The tough job of demonstrating causality

1. Typically if you are studying something it is because you think there is a difference (meaningful or statistically significant) between two groups.
2. BUT…**Correlation doesn't imply causality**
3. Good research tries to eliminate alternative explanations by controlling for other possible causes of the results.

Frequently-discussed research topics[2] with varying degrees of relation.

1. Smoking and lung cancer
2. Media violence and aggression
3. Exposure to lead and IQ test scores in children
4. Calcium intake and bone mass
5. Time spent on homework and academic achievement

[2] Bushman, B.J. and Huesmann, L.R. (2001) in Singer, D. and Singer, J. (Eds.)

Research Typology number two: Degree of experimental control

1. **Experiment**
    a. Has greatest amount of experimental control
    b. Involves random assignment to treatment or experimental condition.
    c. Best research design for demonstrating causality.
    d. Cannot be used to address every research question (e.g., smoking and lung cancer)
2. **Quasi-experiment**
    a. Research design without random assignment to treatment or experimental condition.
    b. If well done, can be used in support of causal relationship.
3. **Non-experimental**
    a. Observational.
    b. Ineffective for demonstrating causality.

Research Typology number three : One shot *vs.* Longitudinal *vs*. Cross-sectional

1. Most research is a **one-shot** study
    a. Is a two-hour 'Introduction to Research Design' course an effective way to teach librarians about research design?
    b. While this may be useful it lacks the strength of comparison information.
2. **Longitudinal:** same people/multiple times
    a. More desirable but expensive
    b. Mortality (drop-outs, particularly selective dropping out).
3. **Cross-sectional**: Different people/Same Time
    a. Drop-outs not an issue
    b. Often less-expensive

Research Typology number four: retrospective vs. prospective

1. **Prospective**: from the present into the future
    a. Future of Children project—followed from birth
    b. NAEP ("Nation's Report Card")
2. **Retrospective:** participant required to report on the present and past
    a. Requires participant to recall the past.
    b. Much epidemiological work is retrospective because the person got into the study because they manifested the disease of interest.
    c. "How many epidemiologists does it take to change a light bulb? None if it's a retrospective study, because the light bulb has already changed itself." [3]
    d. Prospective is stronger design than retrospective but not always feasible.

[3] From *www.childrensmercy.org/stats/definitions/retrospective.htm*

Summary of Part I: Variables and Constructs

Typologies of Variables

1. Physical, demographic, achievement, opinion/attitudinal
2. Levels of measurement: nominal, ordinal, interval/ratio

Constructs (concepts) and operational definitions"

- "and how did you measure that"

Summary of Part II: Typologies of research designs

Typologies covered:

1. Goal: Descriptive, Relational, or Causal
2. Degree of Experimental control: Observational, Quasi-Experiment, Experiment
3. Time: Prospective or Retrospective
4. People: Sample or Population
5. Data Collection: Longitudinal or Cross-sectional

Conclusions

- Correlation doesn't imply causality. If you want to demonstrate causality it has to come from the research design.
- Research is only as good as the operational definitions of constructs that serve as proxies for the constructs of interests. Never be afraid to ask "and how did they measure that?"

Part III: Hallmarks of good research

All research falls somewhere on a continuum on each of these characteristics

Hallmark: Applicability/Relevance (not in the information science sense)

1. Have we answered the question we're interested in?
2. A well-designed study that doesn't address the question is not a good use of resources.
    a. For example, "Was SLA's no handout policy a good idea?" Depends what you mean by effective.
3. Will anything be done differently based on how the study turns out?
    a. Will the results be available before the decision needs to be made?
4. Different stakeholders may have different needs .

Hallmark: Ethical

There is a legacy of unethical research[4] that has resulted in the use of Human Subjects Committees. But they are there for a reason.

1. Research should not cause physical or psychological harm.
2. Research requires informed consent
    a. Not always easy to obtain without disclosing the intentions of the study.

The tension between informed consent and naïve participants is typical of the trade-offs required in conducting research.

- A highly ethical study may not answer your research question
- A highly reliable measurement tool may not be a valid measure of what you're trying to measure.
- A highly controlled research setting may not produce highly generalizable results.

[4] Blass (2004), Reverby (2000), Zimbardo (2007),

Hallmark: Generalizability

1. Ideally, you'd study the population of interest. However, this is almost always impractical.
2. A well-designed study should allow the researcher to generalize of results from sample to population we're actually interested in (a.k.a., external validity).
3. Types of measuring tools
    a. Objective measures
        i. Physical properties using ruler, scale, etc.
    b. Less objective measures:
        i. Multiple-choice questions to measure knowledge
        ii. Demographic questions
    c. Least objective measures
        i. Attitudes/opinions using surveys
        ii. Performance using judges
        iii. Essays to measure knowledge

Hallmark: Reliability

1. Reliability of measurement tools

a. Inter-Rater or Inter-Observer Reliability
   Used to assess the degree to which different raters/observers give consistent estimates of the same phenomenon
    i. e.g., judges at a gymnastics competition, observers rating "aggressive" behavior

b. Test-Retest Reliability
   Used to assess the consistency of a measure from one time to another.
    i. Score should be identical if nothing has changed.
    ii. Alternatively, use "parallel forms" instead of giving the same test twice if exposure to the earlier form could affect performance.

2. Reliability of respondent's self-reports

a. The degree to which a person responds identically given their response to what is being measured hasn't changed.
    i. Which of these two wordings would be more reliable?
        a. Did you find today's session valuable? (Yes or No)
        b. Today's session was valuable (Strongly agree, agree, disagree, strongly disagree)
    ii. Why might the second wording be more useful?

Hallmark: VALIDITY

The remainder of the session will be about validity.

- Because we don't know what the true state of affairs is…we don't know how valid our work is (or isn't)
- However, there are exceptions, such as predicting a presidential election.

The classic works on VALIDITY in social research

- Campbell, D.T. and Stanley, J.C. (1963). Experimental and Quasi-Experimental Designs for Research.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation : Design & analysis issues for field settings*.

There are actually several types of validity: We'll discuss each of these in turn.

1. **Statistical conclusion validity**: Is there a relationship between the two variables that is masked by the statistical results or vice versa.
2. **Internal validity:** Given that there is a relationship, would this relationship have existed in the absence of the "treatment."
3. **Construct validity:** how well have the constructs used to establish the relationship been operationally defined.
4. **External validity (generalizability):** validity with which conclusions are drawn about the generalizability of a causal relationship to and across populations of persons, settings, and times.

Statistical conclusion validity

1. Is there truly a relationship between the two variables that is masked by the statistical results? (Type 2 error)
2. Alternatively, is there a statistically significant result when there really wasn't a relationship between the two variables. (Type 1 error)
3. Many design elements can increase the likelihood of obtaining a statistically significant result if there really is one.
    a. Large sample
    b. The reliability of your measures (test-retest, inter-rater reliability)
    c. The consistency of your intervention (e.g., interviewer quality)

Internal Validity: Given that we observe a relationship between two variables, is the relationship causal?

OR, IS THERE A THIRD VARIABLE THAT ACCOUNTS FOR THE OBSERVED RESULTS?

"Was SLA's no handout policy a good idea?"

Stakeholder 1: SLA Headquarters: good idea=reduce paper
Stakeholder 2: Elana Broch: good idea=didn't impair ability of session attendees to learn and retain what I had to say.

There are many possible threats to the internal validity of a study. **All researchers know these exist. A well-designed study tries to rule out possible alternative explanations (third variables).**

Design: pretest –posttest. One-shot. Quasi-experimental (manipulation: handout or not). Sample.

I'd have two conditions: those with handouts and those without.
Suppose I observed the following results

Pretest
Posttest

Need to rule out that attendees without handouts…

- didn't download ahead of time.
- look at their neighbor's handout.
- didn't leave in frustration because they couldn't keep up.
- weren't "resentfully demoralized for receiving less desirable treatment."
- Didn't differ in their knowledge of research design from those with handouts.
- Didn't receive less eye contact (feel connection with instructor) because they didn't have handouts.
- Didn't choose not to take a handout because they have no interest in the topic.
- Weren't angry with me for bringing handouts in spite of SLA's request for no handouts.

Pretest on research design questions. Posttest on research design questions. Comparison group if I gave the course again with handouts.

Pretest posttest comparison can allow for a relatively strong design. It can rule out

Using a control group can greatly increase internal validity

Participant characteristics that can threaten internal validity

Threats to internal validity when a study takes place over time
Trade-off between a meaningful interval of time and the introduction of specific threats to validity such as

History : "the specific events occurring between the first and second measurement, in addition to the experimental variable. (p. 5)

Sept. 11, SAT recentering, curriculum changes

Change in measuring instrument (due to unreliable measure); Ceiling effects
Experimental Mortality (death, or more likely, dropping out of study). People who feel the drug they're taking experimentally isn't working or believe they are in the placebo condition (i.e., control group).

i. **Self report errors may be intentional or not**
   1. **Self-reported GPA more accurate than self-reported SAT [5]**

[5] (Cassady, 2001)

Construct validity: how well the construct has been operationally defined.

1. Inadequately defining the construct (e.g., aggression) has implications for the usefulness of your study.
2. Measuring a construct in only one way can be limiting
3. Common threats to construct validity
    a. Participants guessing--correctly or incorrectly-- what the desired response/behavior is.
    b. Participant wanting to seem more psychologically healthy than they are
    c. Experimenter expectancies (interviewer effect)

External validity (Generalizability)

1. Definition: Degree to which conclusions are drawn about the generalizability of a causal relationship to and across populations of persons, settings, and times.
2. C&S, p. 20: “Patent artificiality of the experimental setting…I’m a guinea pig”
3. Were there characteristics of your study that distinguish them from the population you hope to extend your results to?
    a. Participants (e.g., psych majors, compliant people, self-disclosers)
    b. Experimenter (e.g., interviewer skill)
    c. Research setting (e.g., university lab)
    d. Measurement instruments (e.g., pretest sensitizing)

Summary of Part I: Variables and Constructs

Typologies of Variables

- Physical, demographic, achievement, opinion/attitudinal
- Levels of measurement: nominal, ordinal, interval/ratio

Constructs (concepts) and operational definitions

- Research is only as good as the operational definitions of constructs that serve as proxies for the constructs of interests. Never be afraid to ask "and how did they measure that?"

Summary of Part II: Typologies of research designs

Typologies covered:

- Goal: Descriptive, Relational, or Causal
- Degree of Experimental control: Observational, Quasi-Experiment, Experiment
- Time: Prospective or Retrospective
- People: Sample or Population
- Data Collection: Longitudinal or Cross-sectional

Conclusion

- Correlation doesn't imply causality. If you want to demonstrate causality it has to come from the research design.

Summary of Part III: Hallmarks of good research

- Applicable/Relevant
- Ethical
- Reliable
- Valid: statistical, internal, construct, external
- Generalizable

Conclusion

- All research falls somewhere on a continuum on each of these hallmarks.
- There is a dynamic tension between many aspects of good research design.
- Well-designed research addresses these tensions as best as possible, within the constraints of time, money, and ethics.

This is the quotation I showed at the beginning. Is it more meaningful to you, now? "To ensure valid results without incurring the time and cost required to interview all members of the target population, researchers develop [research] protocols for drawing a sample of individuals that adequately represented the entire population and from which conclusions about the same could be drawn with a known level of reliability."[6]

I hope I've given you an appreciation for the many aspects of validity and how it impacts research.

And a healthy skepticism, whether in regards to a recommendation to take a medication, support additional funding for a educational approach, or introduce a new flavor of soda.

Vocabulary: Experimental Design

Condition
Treatment Effects
Pre and post effects
Placebo effect
RAPT:Random assignment of participants to treatment groups
Experiment
Quasi-experiment
Convenience sample
Control Group
Placebo condition

[6] From **http://orpheus.ucsd.edu/survey/method.htm**

References

Almond, D., & Edlund, L. (2008). Son-biased sex ratios in the 2000 United States Census. *Proceedings of the National Academy of Sciences of the United States of America, 105*(15), 5681.

Babbie, E. (2008). The basics of social research / 4th ed. Belmont, CA : Thomson/Wadsworth.

Blank, R. M. (2008). Presidential address: How to improve poverty measurement in the United States. *Journal of Policy Analysis and Management, 27*(2), 233.

Blass, T. (2004). The man who shocked the world: the life and legacy of Stanley Milgram. New York: Basic Books.

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

Cassady, J. C. (2001). *Self-reported GPA and SAT scores.* ERIC Clearinghouse on Assessment and Evaluation.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation : Design & analysis issues for field settings*. Boston: Houghton Mifflin.

Huesmann, L. R. (2007). The impact of electronic media violence: Scientific theory and research. *Journal of Adolescent Health, 41*(6), S6-S13.

McClure, S. M., Li, J., Tomlin, D., Cypert, K. S., Montague, L. M., & Montague, P. R. (2004). Neural correlates of behavioral preference for culturally familiar drinks. *Neuron, 44*(2), 379-387.

Parascandola, M. (2004). Skepticism, statistical methods, and the cigarette: A historical analysis of a methodological debate. *Perspectives in Biology and Medicine, 47*, 244-261.

Pendergrast, M. (1993). *For God, Country, and Coca-cola : The unauthorized history of the great American soft drink and the company that makes it*. New York: Scribner's .

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research, 28*, 450-461.

Reverby, S. (2000). *Tuskegee's truths : Rethinking the Tuskegee syphilis study*. Chapel Hill, NC: University of North Carolina Press.

Singer, D. G., & Singer, J. L. (2001). *Handbook of children and the media*. Thousand Oaks, CA: Sage Publications.

Zimbardo, P. G. (2007). Revisiting the Stanford prison experiment: A lesson in the power of situation. *Chronicle of Higher Education, 53*(30)