Educational Research Primer

Lauren Margulieux and Amanda Madden

Center for 21st Century Universities

Table of Contents

The purpose of educational research is to better understand how people learn and improve student learning. Typically, this research asks what students think, such as "What is the percentage of students interested in taking courses online?" or assesses how a change in instruction affects learning, such as "Are online lectures are equally effective as face-to-face lectures?". Your research methods will depend heavily on which of these two goals (or both) you are trying to accomplish and many other factors. If you have little experience with human subjects research or you just need a refresher, this primer is intended to help you select the appropriate methods for conducting educational research. To do this, we'll go through six main sections: types of research questions, pre-tests vs. post-tests, non-experiments vs. experiments, types of measurement, analysis, and statistics.

Types of Research Questions

Many educational research questions come from observing something unexpected in the classroom or reading about a new method of instruction. For example, you might find that a student tried a new way of studying that was effective and you want to know if it would be effective for other students. Or you might have read about flipped classrooms and want to know if using that model would improve learning outcomes in your class. Your research methods will depend heavily on what type of research question you have.

There are three main types of research questions: descriptive, relational, and causal. A good educational research question identifies the group that you are studying, such as online students, and the variables that you intend to measure. Variables in educational research are usually related to a behavior, such as time spent interacting with course materials, or a characteristic, such as gender. Variables can have values that are discrete, meaning that the values are completely independent (e.g., religious association), or continuous, meaning that the values are one point on a spectrum (e.g., test scores).

Good research questions are also open-ended. They typically start with "how," "what," or "why," and cannot be answered with a simple yes or no. For example, the answer to, "How do men and women act differently?" provides much more information than, "Do men and women act differently?"

Descriptive

Descriptive questions ask about the characteristics of variables or groups of people. Answers to these questions can describe differences within or between variables, but they cannot make inferences about the relationships among variables. For example, if you were exploring gender differences in study time, you might find that the women in your research *spent* more time studying than men, but with descriptive research you couldn't say that women *are likely* to spend more time studying than men.

Correct Example 1: What is the ethnic composition of students in online classes?

Incorrect Example 1: How likely are members of different ethnic groups to take online or on-campus courses? This is a relational question.

Correct Example 2: How often do male and female students post on the forum?

Incorrect Example 2: What is the relationship between number of forum postings and gender? This is a relational question.

Relational

Relational questions ask about the relationships among variables, but they do not ask about the cause and effect between variables. Relational questions often ask if a change in one variable is related to a change in another variable. For example, a question might ask if the number of lectures missed is related to course grade. People often equate a relationship between variables with a causal relationship, which says that a change in one variable *causes* a change in the other. However, correlation is not causation for reasons that are detailed in the analysis section.

Correct Example 1: What is the relationship between students' gender and academic major?

Incorrect Example 1: How does students' gender affect choice of academic major? This is not possible to scientifically support for reasons you'll see in the "causal" section.

Correct Example 2: How does time spent studying relate to course grade?

Incorrect Example 2: How does changing students' study time affect course grades? This is a causal question.

Causal

Causal questions ask about the cause and effect relationship among variables. These types of questions are the most demanding of the three because they require the researcher to manipulate the variable predicted to cause an effect. Sometimes researchers would like to find a causal relationship between variables, but are unable to manipulate a variable for practical or ethical reasons. For example, you cannot assign students to a religion for practical reasons nor assign people to smoke cigarettes for ethical reasons; therefore, you cannot determine that religion or smoking causes something else.

Correct Example 1: How does lecture medium (live or recorded) affect students' learning?

(Semi) Incorrect Example 1: How does course medium (online or on-campus) affect students' learning? This question would likely not be causal because you usually cannot manipulate whether students take a course online or on-campus. It would be a correct causal question if you assigned students to online or on-campus courses.

Correct Example 2: What are the differences in learning outcomes between providing immediate or delayed feedback on homework?

Incorrect Example 2: What are the differences in learning outcomes between providing grades throughout the semester or providing only a final grade? This question could likely not be tested ethically.

Is/are your research question(s) descriptive, relational, or causal?

- Descriptive
- Relational
- Causal
- Combination, _____

Type of Question	Definition	Example	Explanation
Descriptive	Asks about characteristics of groups/variables	What is the average number of forum posts for each student?	The answer to this question describes a forum behavior
Relational	Asks about relationship among variables	Are men more likely than women to post on forums?	The answer to this question relates the gender of students to a forum behavior
Causal	Asks about cause and effect relationship among variables	How does number of forum posts by the instructor affect the average number of posts by students?	The answer to this question establishes cause and effect between instructor and student behavior

Pre-tests vs. Post-tests

Pre-Post Design

In educational research, we are often trying to measure what students learn. To do this, we need to measure the level at which students start, meaning their prior knowledge, and the level at which students finish after a course or intervention to make claims about what they've learned. The type of design that measures before (**pre-test**) and after (**post-test**) a course or intervention is called a **pre-post design**. This design is good at measuring any sort of change from before the research started to after the research started, such as how students' knowledge differs from the beginning to the end of the course.

Example:

Research question – How do physics students' learning outcomes differ when they complete homework problems in a group instead of by themselves?

Research design – Give students a test at the beginning of the semester, ask students whether they complete problems in a group or by themselves, give students the same test at the end of the semester, compare test performance between those that completed problems individually and those that worked in groups.

Researchers sometimes include multiple post-tests in this type of design. For example, if you wanted to measure prior knowledge before a course, learning at the end of a course, and retention 6-months after the course, then you could administer the same test at those time intervals. Technically, that would be considered a pre-post-post design.

Post-Only Design

If you need to collect data only after the intervention, then you can use a **post-only design**. This design is good at measuring student attitudes or behaviors that develop over a course but are not present at the beginning. For example, the course evaluation surveys that students complete at the end of the semester are a post-only design. This design is appropriate because students likely don't have a strong opinion about the course at the beginning of the semester. Like the pre-post design, this type of design can include multiple post-tests to create post-post designs.

> Example:
>
> Research question – How do composition students' discussions differ when they use online forums instead of in-class discussion?
>
> Research design – Assign students to use online forums or come to class discussions, measure their discussions (e.g., number of contributions per person) near the end of the course, compare performance between those that used online forums and those that discussed in class.

Note: If you measured discussions at the beginning of the semester as well as near the end of the semester, then you could either have a pre-post or post-post design. If the first measurement for both groups occurred during a classroom discussion (i.e., the conventional medium), then it would be a pre-post design. If the first measurement occurred in the assigned medium (i.e., online for the online group), then it would be a post-post design because the intervention has already occurred.

	Definition	Example	Explanation
Pre-post design	Takes measurements before and during/after intervention to capture change	Give the same test at the beginning (pre-test) and end (post-test) of a course	By comparing the pre- and post-tests, the learning gains can be determined
Post-test	Takes measurements during/after intervention to capture outcomes	Give a final exam (post-test) for a course	Performance on the final exam demonstrates what students know at the end of the course
Multiple post-tests	Takes measurements at multiple points during/after intervention	Give the same test at the beginning (pre-test), middle (post-test), and end (post-test) of a course	Multiple post-tests allow researchers to track progress throughout the course

Based on your research question, do you need a pre-test (i.e., are you measuring a change from before an intervention to after an intervention)?

- Yes, I need a pre-post design

- No, I need a post-only design

Based on your research question, do you need multiple post-tests (i.e., are you measuring a change at multiple points after an intervention, such as for retention)?

- Yes, I need more than one post-test
- No, I need only one post-test

Non-Experimental vs. Experimental Design

The type of research design that you need depends on the type of research question that you have. Descriptive and relational questions can be answered with non-experimental designs, and causal questions must be answered by experimental designs. Note: these design categories are independent from pre-test and post-test designs, so you can have a pre-post non-experimental design or a pre-post experimental design.

Non-Experimental Design (descriptive and relational questions)

In **non-experimental designs**, researchers are measuring phenomena as they exist in the world, and they are not systematically manipulating anything, meaning there is no intervention. Because no systematic manipulation occurs, these designs can answer only descriptive or relational questions. Interactions between researchers and the participants in the study should be limited to what is necessary for collecting data. To collect data, researchers might ask participants to fill out surveys or another type of measure. If direct interaction with participants is impossible or might invalidate the data by biasing participants, an observational approach might be appropriate. In **observational** research, researchers do not directly interact with participants, but they collect data by carefully observing participant behaviors. An example of observational research would be counting the number of contributions from each student in an in-class discussion.

Experimental Design (causal questions)

In **experimental designs**, researchers systematically manipulate a variable to measure how the intervention affects another variable. For education, usually the way a topic is taught is manipulated and learning is measured. This manipulation allows researchers to answer causal questions – it allows researchers to say that they systematically changed one variable; therefore, differences in the measured variable are likely due to that change (the reason "likely" is used will be explained in the analysis section).

If you manipulate some variables and not others, then you have a **quasi-experimental design**. Because some of the independent variables are manipulated (e.g., instructional style) and some are not (e.g., gender), it cannot be a true experiment. Quasi-experimental designs use many of the same methods and analyses as experimental designs. The conclusions drawn from these analyses, though, are different than from experimental designs. Causal relationships can only be concluded if the variable is manipulated. Otherwise we can only discuss the relationship between variables.

Dependent and Independent Variables

As discussed earlier, variables are anything that can have different values. **Dependent variables** are the variables that you collect data about in research, like learning outcomes. They apply to all research

designs: non-experimental and experimental. All of your measurements, such as test scores or attitudes, are dependent variables. These will be discussed more in the measurement section.

Independent variables represent differences in groups that you think might impact the dependent variables. For example, you might hypothesize that an instructional style (either lectures or active learning) affects learning outcomes. In this example, instructional style is the independent variable that differs, and learning outcomes is the dependent variable that you are measuring.

We describe experimental designs primarily by the independent variables. If an experiment has one independent variable, then it is called a **one-way** design. If it has two independent variables, then it is called a **two-way** design, and so on. If you had one independent variable that had four levels (i.e., each student could be assigned to one of four groups), it would be called a one-way design with four levels. For more than one independent variable, we describe them as a ___ x ___ design. The number of blanks is the number of independent variables you have. In each blank is the number of levels for the independent variable. For example, for a study with two independent variables, one with two levels and the other with three levels, it would be described as a 2 x 3 design (read as "two by three"). For a study with three independent variables, each with two levels, it would be a 2 x 2 x 2 design.

Included in these design descriptions is whether participants are exposed to multiple levels of an independent variable. For example, if your independent variable was teaching method (lecture vs. active learning) and if you taught one unit of a course with lectures and another unit with active learning, then the participants (students) would be exposed to both levels of the independent variable. If participants are exposed to all levels of the independent variables, then it is a **within-subjects** design. If you taught one group of students with lecture and another group of students with active learning (e.g., two sections of a course) and each group was only exposed to one level of the independent variable, then it is a **between-subjects** design. You can also have **mixed-subjects** designs in which the participants are exposed to multiple levels of some independent variables but not all of them. Mixed-subject designs are common when one independent variable is within-subjects and another is between-subjects. For example, say you wanted to know if gender affected efficacy of teaching method, and you taught one unit of a course with lectures and another unit with active learning, this design would be mixed-subjects Gender would be a between-subjects variable because each student has one gender, and teaching method would be a within-subjects variable because each student gets both methods.

Independent variables can be **fixed**, meaning they are manipulated by the researcher, or **random**, meaning they are pre-determined. Fixed independent variables (e.g., instructional style) are used in experimental designs, and participants must be able to be assigned to one value of the fixed variable (e.g., receives lecture or active learning). The researcher manipulates the fixed variable to explore its effect on the dependent variable(s). Random independent variables (e.g., gender or religion) are used in non-experimental designs. These variables cannot be manipulated, but they can still represent a difference between groups on dependent variable(s). If you have both fixed and random variables, then you have a quasi-experimental design.

Based on your research question, is your design

- Non-experimental
- Non-experimental and observational
- Experimental

- Quasi-experimental

Variables	Definition	Example	Type of Question
Dependent Variable (DV)	Variable about which data are collected	Test grades, number of forum posts, opinions about online learning	DVs are required for all types of research questions
Random Independent Variable (IV)	Variable of the participant that cannot be manipulated by a researcher	Gender, age, number of courses taken online previously	Random IVs are commonly used in nonexperimental designs to answer descriptive and relational questions
Fixed Independent Variable (IV)	Variable of the participant that the researcher manipulates	The type of instruction (lecture vs. active) that students receive	Fixed IVs are necessary for experimental designs to answer causal questions

* quasi-experimental designs include both fixed and random independent variables

List each of your independent variables (usually there are 1-3), list the levels of each, whether they are between-subjects, within-subjects, or mixed, and indicate whether they are fixed or random. You don't need to include participant characteristics (e.g., gender, ethnicity, religion) unless your research question explicitly includes them. You may still collect information about these differences, but they will not be treated as independent variables in your analyses.

Independent variable	Levels	Between/Within/Mixed	Fixed/Random

Types of Measurement

Operationalizing Variables

Your measurements are your dependent variables. In educational research, one of the dependent variables will almost always be learning. In order to measure learning, we have to define exactly what we mean. Learning could be measured by grades on assignments, such as exams or projects, performance on standardized tests, such as concept inventories, or self-report, such as feelings of learning. All of these options are possible, though some are more defensible in scientific research (more on validity later). As a researcher, you need to **operationalize**, or decide and clearly articulate in a way that can be applied to your research, what you mean when you say learning, or any of your other dependent variables. You might need to operationalize your independent variables too. For example, instead of saying you'll measure "peer-to-peer interaction," say "number of posts on a peer-to-peer forum and number of contributions during peer-to-peer discussions in class."

Types of Data

Your measurements provide the data that you will analyze. There are two main types of data.

Quantitative data represent the world with numbers that can be statistically analyzed. They are necessary for relational or causal research questions. For example, quantitative data could tell us the average number of forum posts per student or the number of times that students watched a video. Quantitative measures are appropriate when you want to confirm a hypothesis (e.g., that students in one group outperform those in another), but they tend to be close-ended, which does not allow for exploration.

Common mistake: Many people accidentally use "data" as a singular noun and will incorrectly say "the data is/shows…" The word "data" is the plural version of "datum," so be sure that you use it as a

Qualitative data are more open-ended than quantitative data and used mainly for descriptive research questions. For example, qualitative data could tell us what a student posts on a forum or what notes a student took while watching a video. Qualitative measures are appropriate when you want to explore a phenomenon (e.g., how students use forums), but they are too detailed and time consuming to amass a large amount of evidence to strongly support a hypothesis. Qualitative data can be quantified by using **coding schemes** that turn descriptions into numbers. For example, for coding a forum in a physics class, you could use a coding scheme that counted the number of times students mentioned each of Newton's Laws of Motion. Quantifying data allows qualitative data to be used in statistical analyses and for relational and causal questions.

Levels of Measurement

Levels of measurement describe the type of quantitative data that you have by categorizing the relationships among values of a variable. Higher numbers do not always mean higher value in data. For example, if you're recording learners' race, you might code "Caucasian" to be 1, "Hispanic or Latino" to be 2, etc. for purposes of analysis. This coding does not mean that Hispanic or Latino is more valuable than Caucasian, but it is merely a way of distinguishing between the two. On the other hand, if you're measuring learners' test scores, then a score of 80 would have a higher value than a score of 70. Levels of measurement categorize these relationships to determine which statistical tests are appropriate to analyze your data. There are four levels of measurement.

Nominal – Lowest Level of Measurement

For **nominal** data, you are basically replacing the name of a value with a number. Like in the race example from earlier, the number does not imply anything about the relationship between values. For another example, if you separated students into groups, the group number would not provide any information about the value of the group.

Ordinal

For **ordinal** data, you can rank-order the values, but the distance between values is not meaningful. For example, you could code prior education as high school degree = 1, some college = 2, college degree = 3, etc. You could rank these values from less education to more education, but the difference between 1 and 2 isn't necessarily the same as the difference between 2 and 3.

Interval

For **interval** data, the difference between values is meaningful. For example, if learners rate how much they liked an activity on a scale of "0 – Not at all" to "10 – A great deal," then the difference between 4 and 5 is equal to the difference between 5 and 6. For interval data, zero is just another number on the scale; it does not indicate an absence of something. In this example, the scale could just as easily start at "1 – Not at all," and the meaning would be the same.

Ratio – Highest Level of Measurement

For **ratio** data, the difference between values is meaningful and zero indicates an absolute zero, or a lack of something. For example, grades are ratio because the difference between 70 and 80 is the same as the difference between an 80 and 90, and a grade of 0 means that nothing about the topic was known (or demonstrated to be known).

Typically, you want the highest level of measurement that makes sense for the data. For example, you'd rather have numeric grade values, which are ratio, than letter grade values, which are interval or arguably ordinal. It would not make sense, however, to measure years in school, which is ratio, instead of highest level of degree earned, which is ordinal, because that level of detail would not be meaningful. The higher that your level of measurement is, the less restricted and more sensitive your statistical analyses can be. That being said, there are few differences between analyzing interval and ratio data, so if your data don't have an absolute zero point, that's not a problem.

Level of Measurement	Definition	Example	Explanation
Nominal	Numbers are placeholders for categories	0 = male 1 = female	Data don't provide information about relationship between values
Ordinal	Numbers provide rank-order but not exact difference between categories	1 = low 2 = medium 3 = high	Data provide information about rank but not exact differences
Interval	Numbers provide information about difference between categories	1 = Strongly Disagree 2 = Disagree 3 = Neutral 4 = Agree 5 = Strongly Agree	The difference between 1 and 2 are equal to the difference between 2 and 3
Ratio	Numbers provide information about difference between categories and zero means an absence	0 = 0 forum posts 1 = 1-5 forum posts 2 = 6-10 forum posts 3 = 11-15 forum posts 4 = 16+ forum posts	The difference between 1 and 2 are equal to the difference between 2 and 3, and 0 means no posts were made

Demographic Data

Demographic data are used to describe relevant characteristics of the participants in your research. We collect demographic data to justify that our sample is representative of the population we are targeting. We might also run correlational analyses (in analysis section) between dependent measures and demographic data to see if participant characteristics are affecting the results. Any distinguishing characteristics of or within the population should be measured in the demographics. These data typically aren't used heavily in the analysis, but are mainly used to describe the sample. Table 1 gives examples of common demographic questions.

Table 1

Questions	Level of Measurement	Response Options
What is your gender?	Nominal	Male, Female, Other
What is your age?	Ratio	_____
What is your employment status?	Nominal	Employed for wages Self-employed Unemployed Homemaker Student Retired Unable to work
What is the highest level of education that you have completed?	Ordinal	Some high school High school diploma GED Some college Associate's degree Bachelor's degree Some graduate school Master's degree Professional degree (M.D., J.D., Pharm.D., etc.) Doctoral degree (Ph.D., T.D., etc.)
Are you a domestic or international student?	Nominal	American (domestic), International
How do you describe yourself? (pick one or check all that apply)	Nominal	American Indian Alaska Native Pacific Islander East Asian South Asian Black Hispanic or Latino Non-Hispanic White or Caucasian
What is your major?	Nominal	_____
Please report your high school GPA if you remember it	Interval	____ out of _____ (e.g., 4.0)

Which institute are you currently attending?	Nominal	_____ or list of choices
Which best describes how many years you've been in college?	Ratio	First-year Second-year Third-year Fourth-year Fifth-year Other
Please report your college GPA if you know it	Interval	_____ out of 4.0
Have you taken any (domain) courses in high school and/or college? (might split into two questions if warranted)	Nominal	Yes, No
If so, how many?	Ratio	_____
What were their names?	Nominal	_____
Do you consider English to be your primary language?	Nominal	Yes, No
If not, what is your primary language?	Nominal	_____
How comfortable do you feel using a computer?	Interval	1 – Not comfortable at all 2 3 4 – Neutral 5 6 7 – Very comfortable
How difficult do you think learning to (course outcome) will be?	Interval	1 – Very difficult 2 3 4 – Neutral 5 6 7 – Very easy

Survey Research

Surveys are prominent in educational research for measuring students' opinions. Exams can even be considered surveys to measure students' achievements. Data collected from survey research can be any level of measurement. For example, if you ask the question "Did you like this instructor?" you could collect ordinal data (e.g., "no," "some," or "yes") or interval data (e.g., "1 – very little," "2," "3 – some," "4," "5 – a lot").

Interval data is commonly collected with Likert-type scales. The classic Likert (pronounced lick-ert) scale is a 5-point scale as depicted below.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1	2	3	4	5

Likert-type scales can range from 3 to 7 points, depending on how much sensitivity is desired. People tend to be less reliable when making more than 7 distinctions, so providing more choices can lead to unreliability. If you want to force people to choose an option other than neutral, provide an even

number of choices to avoid a neutral option. The anchors/endpoints for Likert-type scales can be anything, but people are most familiar with "Strongly Disagree" to "Strongly Agree."

There are plenty of easily accessible resources on the internet to help design better surveys, but we'll include a few tips below.

Tips for Writing Effective Surveys	Examples of Bad Questions	Better Alternatives
1. Don't use one question to ask more than one thing.	Do you think the instructor is smart and likeable?	Do you think the instructor is smart? Do you think the instructor is likeable?
2. Don't use jargon or complicated sentence structure.	How often did you receive formative feedback?	How often did the instructor give you feedback while you worked on assignments?
3. Don't use ambiguous questions.	Did you have several meetings with your group?	Did you have more than 5 meetings with your group?
4. Don't overlap answers.	What is your age? 20-25, 25-30, 30-35, etc.	What is your age? 20-24, 25-29, 30-34, etc.
5. Don't use leading language.	Should the instructor spend more one-on-one time with students?	Were you satisfied with the amount of one-on-one time spent with students?
6. Don't use strong language.	Was the instructor fantastic?	Was the instructor competent?
7. Don't use false dichotomies.	Do you think the instructor is smart or likeable?	Do you think the instructor is smart? Do you think the instructor is likeable?
8. Be selective about which questions to include. Long surveys are abandoned surveys.		
9. Include a non-committal option such as "N/A" or "Prefer not to answer."		

A Quick Note on Validity and Reliability

When we talk about **validity** in research measurement, we're talking about the extent to which the measurement we use actually measures what we think it measures. When we talk about **reliability** in research measurements, we're talking about the extent to which the same participant would give you the same score on a measurement that was administered more than once. In other words, validity and reliability question whether your measures will be legitimate and consistent, respectively. If you have low validity or reliability, then you'll have a lot of error in your data and be more likely to make errors in your conclusions.

Discussing the four types of validity (internal, external, construct, and conclusion) and four types of reliability (inter-rater, test-retest, parallel-forms, and internal consistency) is outside the scope of this primer, but you'll generally pass validity and reliability standards if you let prior research be your guide. For example, in your research area if there's a typical way of measuring and analyzing learning

outcomes, it's likely because it's been deemed valid and reliable by the scientific community. Your findings will be much more convincing if your data are collected using the conventions in your area (e.g., using concept inventories instead of tests that you made). There are statistical methods for demonstrating validity and reliability such as principle components, factor analysis, Cronbach's alpha, etc., but many of them tend to be subjective or require large amounts of data. More information about reliability and validity can be found in any human subjects research methods textbook.

List your dependent variables and indicate which level of measurement they are. Note: if you have a survey or test as a dependent variable, you do not need to list each individual question separately.

Dependent variable	Level of Measurement

Analysis

Creating a Data File

When you create your data file, analytic software expects that individual participants are represented on the rows and your variables are represented on the columns like in the example table below.

Participant #	Independent variables		Demographic Data		Measurements	
	IV1	IV2	Age	Major	DV1	DV2
1	1	0	21	3	19.5	5
2	0	1	20	2	20	4
3	1	1	20	1	17.5	6
…	0	0	19	4	15	6

To use statistical analysis software, your data will need to be numeric, though it can be at any level of measurement. Be sure that you keep a codebook to link your numeric codes to their actual meaning. This is imperative so that you don't forget what your codes are and so other people can decipher the data file if needed. For example, for the demographic data under "Major," a codebook would tell you what "3" means.

Analytics Software

SPSS and Stata

Pros: Best for straightforward statistical analyses. SPSS and Stata are easy to learn. Includes support and vets new statistical techniques. Can handle big data.

Cons: Licensing is expensive. Does not support abnormal sets of data or complex analyses.

SAS

Pros: Supports greater complexity in statistical analyses and includes more advanced statistical techniques than SPSS and Stata but is slightly harder to learn. Includes support and vets new statistical techniques. Can handle big data.

Cons: Licensing is expensive.

R

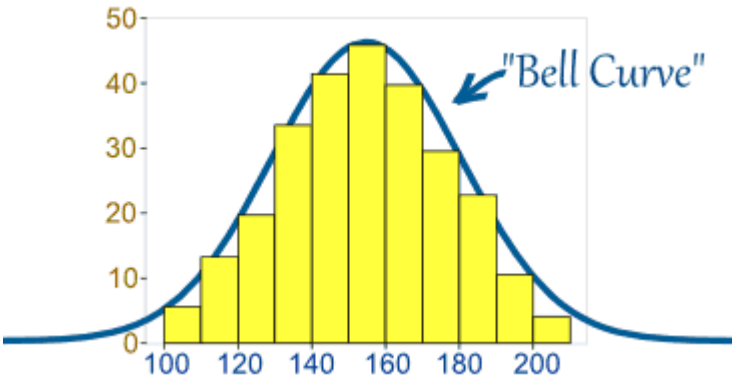
Pros: Free analysis software that can do everything the others can do. It is open source, so the latest techniques are released in R first.

Cons: All analyses must be written in code by hand. There is not interface to help you set up your statistical tests, so it has a very steep learning curve. Doesn't handle big data well.

Distribution of Scores

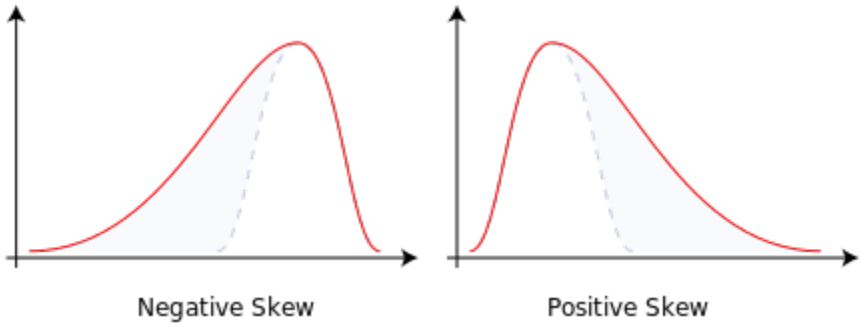
Once you have all of your data, you'll want to look at a histogram of the frequency all of the scores for each dependent variable. In the histogram, the y-axis will be frequency, and the x-axis will be scores on the dependent variable. You're looking to see if your scores are **normally distributed**. Normal distributions follow a bell curve (see figure below). Hint: If you have more than 5-10 possible scores on the dependent variable, your histogram will likely have big gaps in it. You might bundle some groups together to balance out these gaps. For example, if you gave a test on which the possible scores were 0-100, you might bundle scores that are A, A-, B+, B, etc. Otherwise you might see a gap for, say, 86 that could make the distribution look non-normal.

Don't be discouraged: Real data won't fit the normal distribution as perfectly as in the figure. As long as the general shape of your data is similar to the bell curve, you can assume a normal distribution.



Most statistical tests are intended to be used with data that are normally distributed. Distributions can deviate from the normal distribution in kurtosis or skewness. Kurtosis, meaning how flat or tall the distribution is, largely doesn't matter for basic statistical tests as long as the distribution is symmetrical. If your data are skewed one direction or the other (see figure below), then you'll need to be careful about which analyses you use. Data can also be multi-modal meaning that it has more than one peak. Multi-modal distributions cannot be analyzed with the statistical tests described in the "Statistics"

section. Non-normal distributions are typically analyzed with non-parametric tests such as Mann-Whitney-Wilcoxson or Kruskall-Wallis.



Error versus Effect

In 99.9% of human subjects research, you are collecting data to support or refute hypotheses, not to prove or disprove them. We frame research in this way because we use samples of populations rather than whole populations. A **population** is the group of people that you are interested in studying. For example, a population might be the students in a particular major, those at a particular university, in a particular country, or all university students. In most cases, we aren't able, nor is it necessary, to include everyone in a population in a study, so we rely on **samples**, or representative subsets, from those populations. For example, if you were interested in a project about improving physics education (i.e., population of physics students), you might use students in an intro physics class as your sample. Sampling introduces potential error into the research because samples can differ (e.g., students in one physics class aren't the same as those in another), whereas a population is all inclusive.

We also do not attempt to prove hypotheses because human subjects research includes error. **Error** is anything that might influence the results of a study in a way that is inconsistent or that isn't being measured. For example, personal issues may or may not affect a student's performance in a course. Unless the study is about coping with personal issues, this is likely to be a source of error and impact the results of the research. People are incredibly complex, and human subjects research inherently has a lot of error because people's performance in a study can be influenced by almost anything including time of day, day of week, our personality, events that happened yesterday, events that happened a month ago, what we think about the researcher, and what we think the research is about.

All measurements include some degree of error as well. For example, if you ask participants for their age, you can have two participants who are "25" with almost a year difference in age unless you measure with more specificity, which probably would not be worth your time. Other measurements used in educational research are no different.

To manage this error, human subjects research employs **null hypothesis significance testing** to determine whether a phenomenon is likely due to error or not. Basically what that means is that we use statistical analyses to determine if a phenomenon is less than 5% likely to be due to error, reported as $p < .05$. Said another way, $p < .05$ means that we have 95% confidence that the result is due to differences in the independent variable, not just chance. Results that meet the $p < .05$ standard, are considered evidence, but not proof, that a phenomenon exists. More sophisticated statistical analyses, called effect sizes, can tell you how strong or weak a phenomenon is, but they fall outside of the null hypothesis significance testing paradigm.

Differences that exceed the $p < .05$ standard are considered **statistically significant**. When describing results, it is correct to use the full phrase "statistically significant" instead of only "significant," because significant is used commonly enough in the English language that people might infer the colloquial meaning. The opposite of statistical significance is "statistically nonsignificant," not "insignificant." It is important that you don't confuse *statistical* significance with significance. Results can be statistically significant but not meaningful for a number of reasons. For example, if you're analyzing a MOOC with very large sample size, then a 3% difference between groups could be statistically significant, but a 3% difference likely isn't very meaningful in the big picture. Similarly, results can be statistically nonsignificant but meaningful (i.e., significant). For example, if you expected online and in-class groups of students to perform differently, and they did not, then nonsignificant results are meaningful.

Term	Definition	Relationship to Error
Sample	Subset of people from the population that you are trying to reach	Samples include error because they do not necessarily represent all people in the population of interest equally
Normal Distribution	Expected distribution of scores based on inherent differences among people	Because people vary on innumerable characteristics, some differences between participants is expected
Statistical Significance	Conclusion drawn from null hypothesis significance testing that the difference between groups is greater than what could be expected due to chance	To have statistical significance, the difference between groups must be larger than the error within groups.

Statistics

Descriptive Statistics (descriptive questions)

Descriptive statistics provide summaries of your data. They are intended to describe the data as they are rather than draw conclusions beyond your sample of participants; therefore, they are used for descriptive questions. For example, if you had two groups with average test scores of 85 and 88, you could use descriptive statistics to say that one group scored higher than the other. This difference does not necessarily mean that there is a statistically significant difference between the two groups (i.e., that the difference is likely not due to chance), and you could not claim that you'd expect to see the same difference in other classes. You would need inferential statistics to determine if the difference is statistically significant (discussed below) and could be expected in other classes.

Mean

In educational research, we are typically comparing groups of students to other students, especially for quantitative measurements. In this type of paradigm, the unit of measurement of interest is a group's score rather than an individual's score. The mean (*M*), or average score, is typically the most appropriate descriptive statistic to represent a group with a normal distribution. If you have skewed data, then you might consider using the median, or middle score, instead.

Standard Deviation

Besides the mean, we typically also want to know how homogenous a group's scores are. For example, a group that has an average test score of 85 with all individual scores in the 80s is more homogenous than a group that has an average test score of 85 with individual scores between 70 and 100. **Standard deviation** (*SD*) is a unit that allows you to describe how dispersed scores in a group are. In the test score example, the first group's scores are closer together, so the standard deviation would likely be less than in the second group.

If data follow a normal distribution, about 68% of scores will fall within one standard deviation of the mean, 95% of scores within two standard deviations, and 99% within three standard deviations. For example, if a group has a mean of 85 and standard deviation of 2 (and a normal distribution), then 68% of scores will be between 83 and 87, 95% of scores between 81 and 89, and 99% of scores between 79 and 91.

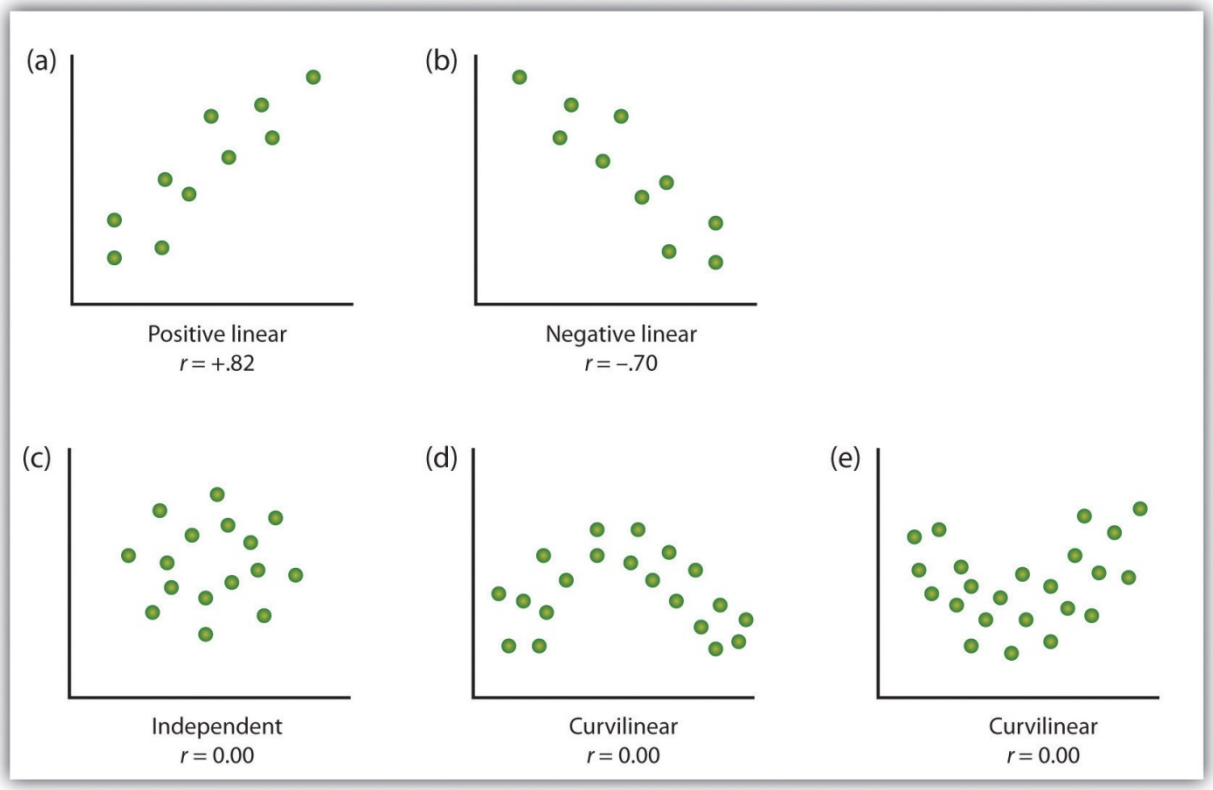
Inferential Statistics (relational and causal questions)

The purpose of **inferential statistical analysis** is to determine if the differences between groups is greater than what would be expected due to chance. Inferential statistics allow you to make some generalizations about your findings beyond your sample. For example, if you tested a new teaching method and found that students who received this new teaching method scored 5 points higher on a test than students who didn't, it would matter if the standard deviation were 2 points or 10 points. If the standard deviation were 2 points, then the mean of the intervention group is 2.5 standard deviations higher than the other group, and the intervention very likely had a positive effect. If the standard deviation were 10 points, then the difference in means between the groups isn't greater than what would be expected due to chance.

Correlational Analyses (relational questions)

For relational questions, you are trying to determine the type of relationship between two variables. A **positive relationship** suggests that as the value of a variable increases or decreases, the value of the other variable also increases or decreases, respectively. For example, as attendance in class increase, scores on a test increase. A **negative relationship** is the opposite and suggests that as the value of a variable increases or decreases, the value of the other variable decreases or increases, respectively. For example, as absences from class increase, scores on a test decrease.

These are both types of linear relationships; the relationship is constant across different level of the variables. There are also curvilinear relationships in which the relationship changes across different levels of the variables. For example, as students' level of stress about an exam increases, their scores might increase up to a point, but at that point, excess stress might negatively relate to test scores. These types of curvilinear relationships cannot be analyzed with correlations. A scatterplot of the variables will show you the type of relationship that you have. The images below provide examples of what each type of relationship looks like in a scatterplot. Your data likely won't be as neat, so you'll need to ignore the random points that don't follow the general relationship.



Correlation vs. Causation

For correlational analyses, a relationship should not be misinterpreted as causation for two reasons: the third variable problem and the directionality problem. The **third variable problem** states that a third variable, separate from those in the correlation, might be mediating the relationship. In the stress and test score example from earlier, it is unlikely that stress primarily causes better test scores. More likely, stress causes more or better studying, which causes better test scores. The **directionality problem** states that it is unclear from a correlational analysis which variable is causing which. In the stress and test score example, it is unclear whether high stress hurts test scores or poor test scores cause high stress.

You should choose your correlational analysis based on the whether your variables are continuous (i.e., interval or ratio or discrete (i.e., nominal or ordinal). Discrete variables that are dichotomous, a variable that has two values (e.g., yes or no, introvert or extravert), can be either truly dichotomous or artificially dichotomous. Truly dichotomous variables have discrete values (e.g., a participant either has or has not taken Calculus before). Artificially dichotomous variables are continuous variables that the experimenter has split into discrete groups. For example, participants can take a personality test that will rank their personality on a spectrum from introverted to extraverted, and the experimenter could pick a score and classify all participants below that score as introverted and those above the score as extraverted. Artificially dichotomizing variables this way is generally not recommended, but there are rare cases when it's appropriate.

- If both variables are continuous, use Pearson's *r* correlation (this is most common and is typically the default option for analysis software).

- If one variable is truly dichotomous and the other is continuous (i.e., interval or nominal), use point biserial correlation.
- If one variable is artificially dichotomous and the other is continuous (i.e., interval or nominal), use biserial correlation.
- If you have discrete data with more than two values (polytomized instead of dichotomized), then you'll have to use some of the more obscure tests, such as tetrachoric, polychoric, polyserial, Spearman's rho, or Kendall's Tau-B.

The results of a correlational analysis, or **correlation coefficient**, will be a *r* value between -1 and 1. Negative numbers indicate a negative relationship, and positive numbers indicate a positive relationship. The closer the number is to 1, the stronger the relationship between variables. The closer the number is to 0, the weaker the relationship between variables. The strength of the relationship indicates how accurately you can predict one variable if you have a value for the other variable. In strong relationships, having values for both variables is somewhat redundant. In weak relationships, the value of one variable provides little information about the other. Many people misinterpret the strength of the relationship as the slope of the relationship, but a correlation coefficient of 1 does not mean that the variables have a 1-to-1 relationship.

Below is a table for the heuristic cut-offs (Cohen, 1988) to describe strengths of relationships, but the meaningfulness of the strength of relationship depends largely on the variables that you are correlating. In educational research especially, correlation coefficients tend to be low because of the large variability in learners' prior knowledge, level of motivation, etc.

Value of correlation coefficient	Strength of relationship
r >.5 or >-.5	Strong
r = .5 to .3 or -.5 to -.3	Moderate
r = .3 to .1 or -.3 to -.1	Weak
r <.1 or <-.1	No relationship

T-Test (for interval or ratio data)

If you are comparing two groups (e.g., a group that gets conventional instructions and a group that gets a new type of instructions), then you'll want to use a **t-test** to analyze the differences between groups.

There are a few different type of t-tests based on how you collected data.

- For comparing groups that are between-subjects (participants in groups are mutually exclusive), use an independent-samples t-test. For example, if you taught one group with lectures and another group with active learning, then you could compare their test performance using an independent samples t-test.
- For comparing groups that are within-subjects (participants are the same in both groups), use a paired-samples t-test. For example, if you taught one topic with lecture and another topic with active learning, and the same students learned both topics, you could compare their performance on questions about each topic using a paired-samples t-test.
- For comparing a group to the mean from literature or another study, use a one-sample t-test (this is not common). For example, if you were comparing scores on a concept inventory from your class to the mean score from the literature, then you could use a one-sample t-test.

T-tests will give you a *t* value, which is basically the ratio of the difference between groups to the error within groups. The larger *t* is, the bigger the difference between groups is compared to the error within each group. For example, if the difference between group means was 10, the *t* value will be larger if the standard deviation were 5 than if it were 15. The *t* value can be positive or negative depending on whether you have a positive or negative relationship between groups. You can determine if that value is statistically significant based on the *p* value (i.e., $p < .05$ is statistically significant), or you can determine how large the difference between groups is using the effect size statistic Cohen's *d*.

Below is a table for the heuristic cut-offs (Cohen, 1988) to describe the size of effects based on Cohen's *d*, but the meaningfulness of the strength of relationship depends largely on the variables that you are correlating.

Value of *d*	Size of effect
$d > .8$	Large effect
$d = .8$ to .5	Medium effect
$d = .5$ to .2	Small effect
$d < .2$	No effect

ANOVA (for interval or ratio data)

If you are comparing more than two groups, then you'll want to use analysis of variance, **ANOVA**. Comparing more than two groups includes comparing three or more groups of one independent variable and/or groups from multiple independent variables. ANOVA will tell you if there is a **main effect** of variable, a difference between groups within an independent variable, and if there is an **interaction** between independent variables, a difference in an independent variable's effect based on the value of another independent variable. For example, if you had two interventions that you were testing (i.e., 2 x 2 between-subjects design), it could be the case that getting one or the other intervention would not improve test scores (dependent variable). In this scenario, there would be no main effect of either independent variable, meaning that by themselves neither variable improved scores. If students who received both interventions performed better on the test, then that would be example of an interaction. The effect of each intervention relies on the other intervention being given.

The type of ANOVA that you use will depend on whether you have a between- or within-subjects design and the relationships among dependent variables. A standard ANOVA assumes between-subjects design and affords one dependent variable to be analyzed at a time. If you have a within-subjects or mixed design or you expect dependent variable to be related, then you'll need to use one of the other types of ANOVA mentioned below.

ANOVA will give you an *F* value, which is basically the ratio of the differences among groups to the error within groups. The larger *F* is, the bigger the difference among groups is compared to the error within groups. For example, if the group means were 4, 6, 8, and 10, the *F* value will be larger if the standard deviation were 1 than if it were 3. The *F* value can only be positive. You can determine if that value is statistically significant based on the *p* value (i.e., $p < .05$ is statistically significant), or you can determine how large the effect is using the effect size statistic *f*. Below is a table for the heuristic cut-offs (Cohen, 1988) to describe the size of effects based on Cohen's *f*, but the meaningfulness of the strength of relationship depends largely on the variables that you are correlating.

Value of *f*	Size of effect
f > .4	Large effect
f = .4 to .25	Medium effect
f = .25 to .1	Small effect
f < .1	No effect

Post-hoc analyses

An ANOVA will tell you whether there is a difference among your groups, but because you are comparing more than two groups, it will not tell you between which groups the difference occurs. For example, if you were comparing 3 groups and found a statistically significant *F* value, it could be the case that groups 1 & 2 are equal but group 3 is different or it could be the case that groups 2 & 3 are equal and 1 is different. To determine the specific pattern of results within an ANOVA, you'll need to conduct post-hoc tests.

For distinguishing between more than two levels within one independent variable, like in the previous example, the most popular post-hoc test is Tukey's Honestly Significant Different (HSD) test because it is conservative and, therefore, doesn't invite skepticism. This test conducts pairwise comparisons (compares the means of two levels of the independent variable) to determine if one mean is statistically larger or smaller than the other. It will give you *q* values, which are like *t* values and represent the differences among groups.

For distinguishing between groups across different independent variables, the most popular post-hoc test is a t-test. For example, if you had two independent variables with two levels each (2x2 design), and you wanted to compare groups that got the same level of one independent variable and different levels of the other independent variable, then you'd use a t-test.

It's generally a good idea in the null hypothesis significance testing framework to use post-hoc tests for only the groups that you expect to be different (either due to your hypotheses or the means of your data). The more post-hoc analyses that you use, the more error you introduce. Because every test has up to 5% error ($p < .05$), each test that you run adds a 5% chance that your results are due to error. To keep error in check, researchers use the **Bonferroni correction** (sometimes called the Bonferroni adjustment). This correction divides the *p* value by the number of tests that you conduct to ensure that the total error in your analyses is no more than 5%. For example, if you conducted 4 t-tests as post-hoc tests for an ANOVA, then you'd need to divide your *p* value by 4, and your results would only be considered statistically significant if *p* were less than .0125.

ANCOVA

ANCOVA stands for analysis of covariance. If you expect that performance on one dependent measure or a demographic measure will be predictive of performance on the dependent measure being analyzed, then you'll want to use ANCOVA to ensure that the manipulation from the independent variable predicts performance on the dependent variable above and beyond those other variables. For example, if you gave participants a pre-test and expected performance on that pre-test to be predictive of performance on the post-test, then you'd want to use ANCOVA to ensure that your independent variable can predict the post-test separate from the pre-test.

Especially for comparing pre-test to post-test scores, many researchers will use gain scores, but it is better to use ANCOVA. Gain scores have fallen out of favor recently because they tend to be unreliable. Gain scores are the educational equivalent of difference scores. Difference scores take two points of data (e.g., a pre- and post-test) that each have an error component and condense them into one data point, which has its own error component. Analyzing difference scores ignores the error components of the original scores, making it less reliable.

MANOVA

MANOVA stands for multivariate analysis of variance. If you expect or design two or more dependent measures to measure the same thing, then you'll want to use MANOVA. For example, if you gave students 3 exams throughout the semester and wanted to use them together as a measure for student learning, then you could use MANOVA to analyze your results. You could also add up all the test scores to make one final score, but similar to the difference scores from the previous paragraph, that would ignore the error components of the original scores and make your analyses less reliable. If you want to test the similarity of performance on two dependent variables, you can use a correlation. If you want to test the similarity of performance on more than two dependent variables, you can use principles components analysis or factor analysis.

Repeated Measures

Repeated measures ANOVA is for research designs that use the same measurements multiple times, usually at different time points. For example, if you gave learners the same test at the middle of the semester, the end of the semester, and 3 months after the semester, you'd use repeated measures to analyze the results. One of the benefits of a within-subjects design like this is that it reduces error by reducing the number of participants and all of the random factors that they bring with them. Repeated measures ANOVA allows the analysis to capitalize on this reduced error, but it assumes that participants completed all of the measures. If you have more than 5-10% of participants who did not complete every measure, then you should not use repeated measures ANOVA.

Both MANOVA and repeated measures ANOVA analyze multiple measurements from the same participants. The major difference between that two is that in MANOVA, the measurements are different but all measure the same construct, and in repeated measures ANOVA, the measurements are the same and administered at different times.

Linear Regression

Regression is an analysis technique for determining how much of the performance on a dependent variable can be predicted by the independent variables, demographic characteristics, and/or performance on other dependent variables. If you're thinking that it sounds a lot like ANOVA, then you are right. ANOVA, however, is typically used for analyses with few predictors, and it is used much more frequently in the null hypothesis significance testing framework. Regressions typically attempt to account for as much of the variance in the performance of the dependent variable as possible, so it is typically used for analyses with several predictors. Regression also focuses on how well each predictor explains the dependent variable rather than if the differences are statistically significant, so it's used outside of null hypothesis significance testing much more frequently. A regression coefficient, or β, refers to how much of the variance in the dependent variable is predicted by a variable (independent, demographic, or dependent).

Statistic	Type of Question	When to Use
Mean	Descriptive	Find the average score of a group
Standard Deviation	Descriptive	Find the average error of a group
Correlation Coefficient	Relational	Determine the strength of the relationship between 2 variables
T-Test	Causal	Determine if difference between groups on dependent variable is caused by independent variable with 2 levels
ANOVA	Causal	Determine if difference among groups on dependent variable(s) is caused by independent variable(s) with 2 or more levels
Regression	Causal	Determine how much of the variance in a dependent variable is attributable to other variables (e.g., demographics, independent variables)

Additional Analyses

Intra-Class Correlation Coefficient

Intra-class correlation coefficient is an analysis of interrater reliability. If you have qualitative data that needs to be scored and if that scoring scheme is at all subjective, then you'll want multiple raters to score at least some of the data to ensure that the raters are being unbiased. If you have more than one rater, then you'll need to determine how similar each rater's scores are. **Interrater reliability** determines how similarly raters scored data.

There are two types of intra-class correlation coefficients. The first, intraclass correlation coefficient of consistency, ICC(C), is used when you are determining if raters put items in the same order. For example, if you were trying to rank your students from the highest to lowest performers, then you could determine how similar rater's rankings were using ICC(C). The other, intraclass correlation coefficient of absolute agreement, ICC(A), is used when you are determining if raters gave each item the same score. For example, if you were giving each student a numerical grade, then you could determine how similar rater's ranking were using ICC(A). As you might suspect, having a high ICC(A) is more difficult than having a high ICC(C), but an acceptable score on initial rankings for either is .80 or higher. Once you have the initial scores from each rater, disagreements among raters are typically resolved through discussion until 100% agreement is reached. If you have an ICC lower than .80 on initial scores, then you'll need to retrain raters and re-score the data.

If you have too much data or too few resources for multiple raters to score all of the data, then you can take a sample of 20% of the participants and ask multiple raters to score only those participants' data. If the reliability for this sample of the data is .80 or higher, then it is acceptable to have one rater score the rest of the data independently.

Demographic Analyses

One of the reasons to collect data about demographic information is to determine if characteristics of the participants affects performance on the dependent variables. It is good practice to run correlations between demographic data and dependent variable data to see if there is in fact a relationship between the two. If you do find that one of your demographics is correlated with a dependent variable, then you'll want to ensure that there are no meaningful differences among groups on that demographic characteristic. Otherwise it'll be difficult to argue that the differences on the dependent variable are due to the independent variable instead of the demographic difference. You can determine whether there are differences among groups by using descriptive statistics or by treating the demographic variable as a dependent variable and using inferential statistics, depending on what is appropriate for your data. Though this isn't how inferential statistics are meant to be used, this analysis will tell you if there are differences among groups.

For more information

The Research Methods Knowledge Base is a free online resource that has much more information about the topics discussed here.

If you are interested, there are many textbooks about human subjects and social science research methods that will give you more information about the topics discussed here. There isn't one definitive book that everyone uses because each book has a slightly different emphasis. For example, some books primarily talk about behavioral research (Leary's Introduction to Behavioral Research Methods) while others focus on survey research (Fowler's Survey Research Methods) and others focus on qualitative analysis (Miles' Qualitative Data Analysis). For more on effect size and power analysis, see the following reference:

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates: Hillsdale, NJ.

To search for education research articles you can use the common general databases, such as Academic Search Complete or Proquest, and ERIC, Educational Resources Information Center. Many of the papers on ERIC are open source, so you do not necessarily have to have a subscription to find valuable articles.

There are several professional societies that focus on education. The most prominent ones are in the table below. UCF's Faculty Center for Teaching and Learning also has a comprehensive list of publications for scholarship of teaching and learning (SoTL).

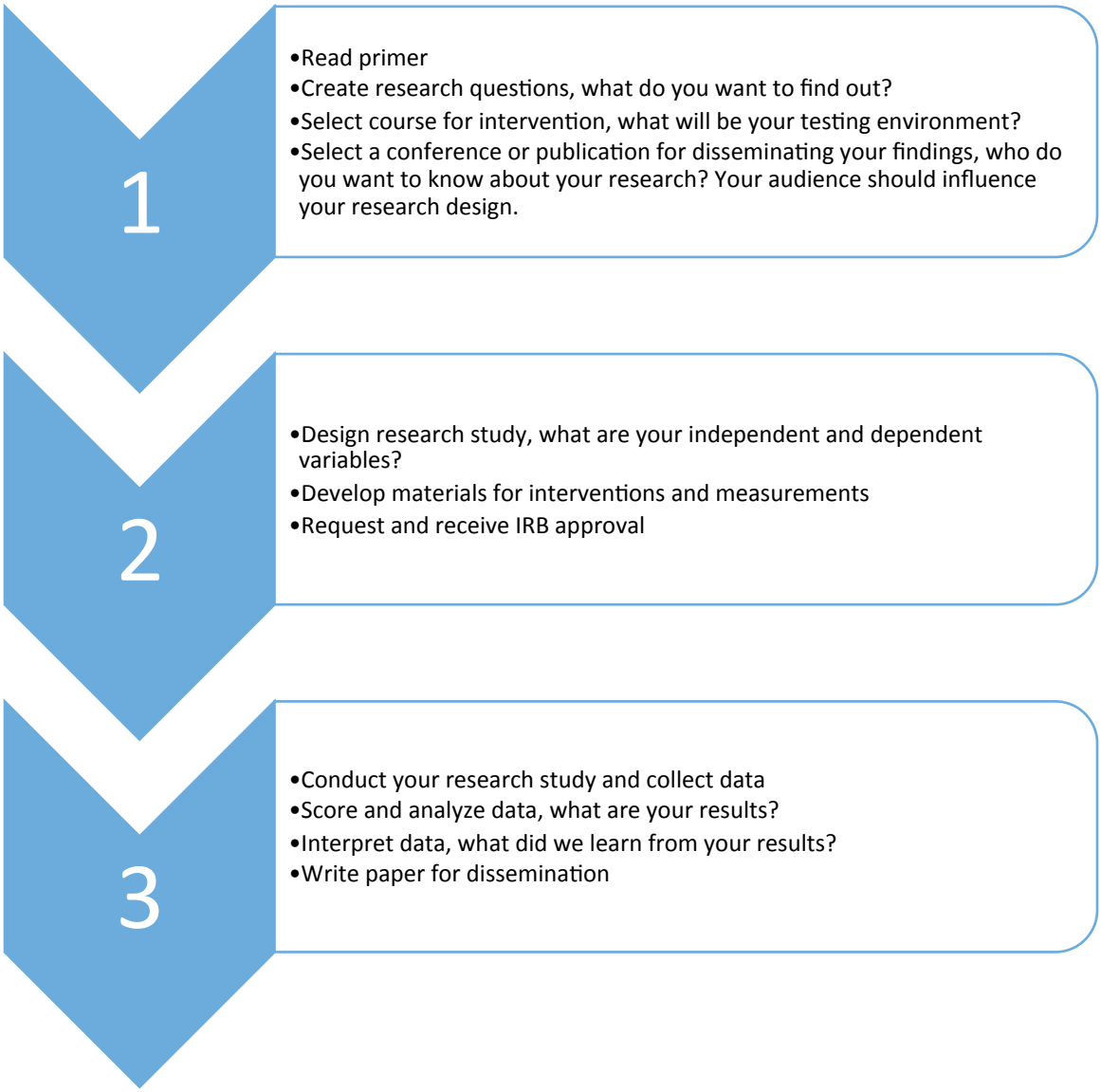
Domain	Society	Conference	Publication
General	American Educational Research Association (AERA)	AERA Annual Meeting	AERA Publications
General	International Society of the Learning Sciences (ISLS)	ISLS Conferences	ISLS Publications
General	APA Division 15 Educational Psychology		Educational Psychologist
Engineering	American Society for Engineering Education	ASEE Annual Conference	ASEE Publications

	(ASEE)		
Computing	ACM Special Interest Group on Computer Science Education (SIGCSE)	SIGCSE Conferences	SIGSCE and related publications

To apply for grants to fund education research, you can start by searching the National Science Foundation's directorate for Education and Human Resources.

Research Process

Sections 1 and 2 of the following process are iterative. You should continually be revising your research questions, design, and materials until you start collecting data.



Glossary

ANOVA

Between-subjects design

Bonferroni correction

Causal question

Correlation coefficient

Demographic data

Dependent variable

Descriptive question

Descriptive statistics

Directionality problem

Error

Experimental design

Fixed independent variable

Independent variable

Inferential statistics

Interaction

Interrater reliability

Interval data

Levels of measurement

Main effect

Mixed design

Negative relationship

Nominal data

Non-experimental design

Normal distribution

Null hypothesis significance testing

Observational research

One-way design

Operationalize

Ordinal data

Population

Positive relationship

Post-only design

Post-test

Pre-post design

Pre-test

Qualitative data

Quantitative data

Quasi-experimental design

Random independent variable

Ratio data

Relational question

Reliability

Sample

Standard deviation

Statistically significant

Third variable problem

t-test

Two-way (or more) design

Validity

Within-subjects design