Research Methods for Psychology

T.L. Brink, Ph.D., M.B.A.
Crafton Hills College

2017

Table of Contents

Chapter #1: Science & Theories

What is Science?

Science is a method of gaining knowledge. Scientific knowledge is based upon *empiricism*, the method of direct, precise, and objective measurement of natural phenomena. Compared to other claims about human knowledge, science is more cautious, more guarded, more reluctant to yield to a false claim, or even yield to a true claim (prematurely).

"Don't believe everything you think."
-Jeremy Berg, Editor-in-Chief of Science Journals

If a phenomenon cannot be studied empirically (e.g., certain spiritual claims resist such investigation) then that phenomenon cannot be studied scientifically.

Each scientific measurement yields a specific datum (fact). These results are known as *data* (the plural form of datum). So, when writing for the sciences, we say “these data were” but in a field such as information technology, where data refers to a big collection of bits, the phrasing would be “this data is.”

The history of science is not just the accumulation of more data. The history of science through Aristotle, Bacon, Descartes, Galileo, Locke, Newton, Darwin, and Einstein is the history of successive approximations of better methods to get better data (and better theories to understand those data).

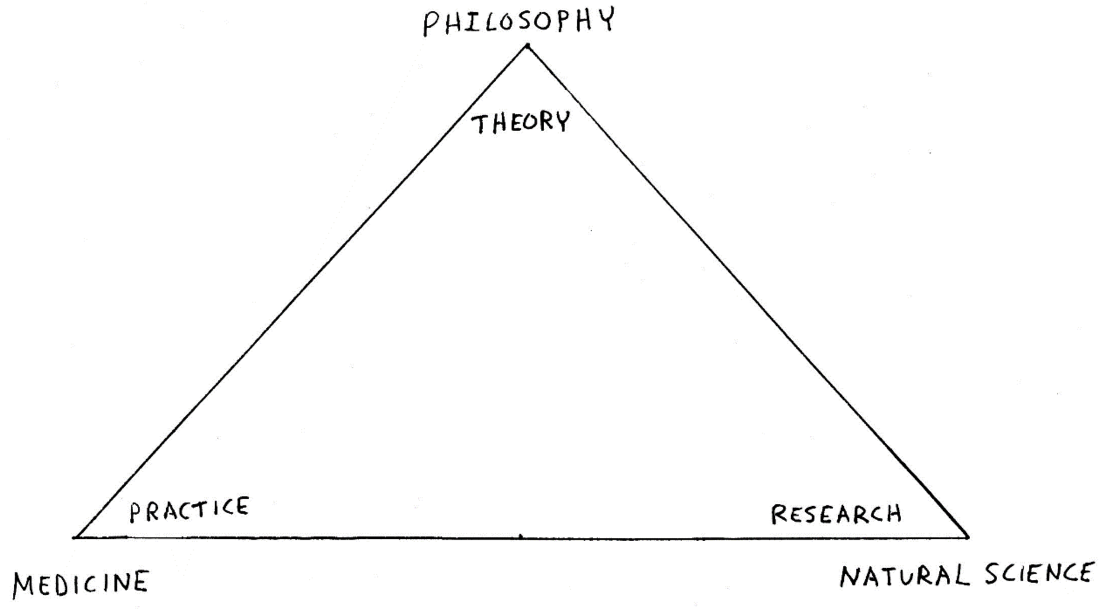
Psychology as a Science

The scientific status of any endeavor is determined by its method of investigation (the empirical method). Science is determined by how something is studied rather than what is being studied. *Psychology* is the scientific study of behavior (and mental processes) in humans and animals. Psychology observes organisms (humans or animals) and tries to understand their *responses* (behaviors) in terms of organismic variables and *stimuli* (external influences). Think of all psychologists as scientists who study behavior.

Psychology is a relatively young science (it certainly has less of a history than does physics or chemistry). Psychology is sometimes called a “hybrid” science in that it was grafted on to other studies of human behavior coming from the natural sciences (e.g., physics, chemistry, biology) as well as other ways of studying human behavior (e.g., medical practice) and the mind (e.g., philosophy and theology).

So, psychologists were not the first to ask questions such as “Why to people think what they think, feel what they feel, and do what they do”? Much of modern psychological science stands on the shoulders of great physicists (e.g., Fechner), medical professionals (e.g., Nightingale), philosophers (e.g., Aristotle) and theologians (e.g., Augustine). Many of the topics now considered within the field of psychology are also being investigated by scientists in bordering fields (e.g., biology, sociology, political science, economics). Indeed, there are no clear cut demarcation lines between the sciences; there are only different (but overlapping) areas of study.

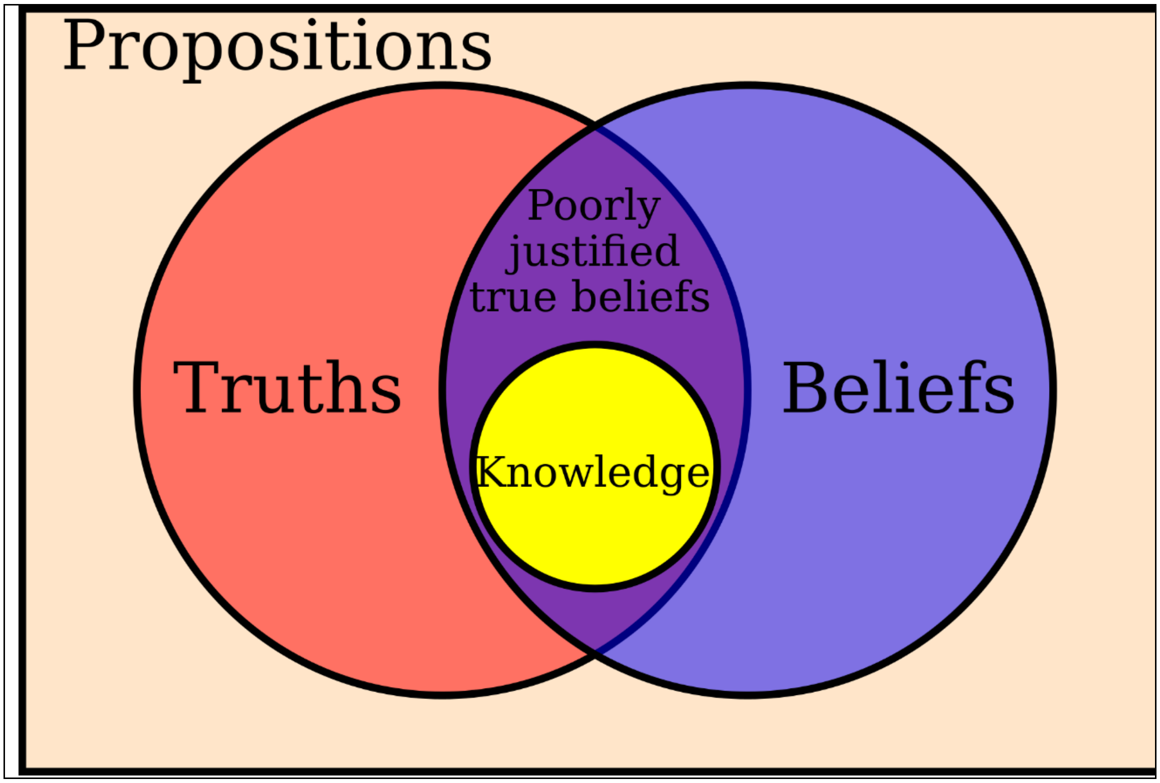
The triangular field of psychology



Today, a good undergraduate curriculum in the field of psychology must cover all the corners of this triangle. Regardless of your future career objective within the field of psychology, you must learn about its theories, research methods, and clinical applications. A course in personality theory or history & systems will cover theory. A course in abnormal psychology will be the one that gets closest to clinical practice. It is the course represented by this textbook that that covers research methods.

Basic vs. Applied Research

Psychologists studying abnormal, consumer or organizational behavior are primarily interested in *applied* research having immediate application answering specific questions about better assessment and intervention. All branches of psychology are interested in *basic* research that gathers data, and can help fine-tune our understanding of behavior, whether we are talking about how fast a rat can learn to run through a maze or how well a psychotherapy patient recovers from depression. There is growing recognition that today's basic research may have I/O, consumer, or clinical applications much sooner (and in ways other) that the original researchers imagined. In the medical field the term *translational* science refers to the fact that research from the "benchside" may soon be applied to the "bedside" of patient care. In the field of education, we sometimes refer to applied research as *action* research, especially if it quickly takes advantage of a serendipitous opportunity to gather data.

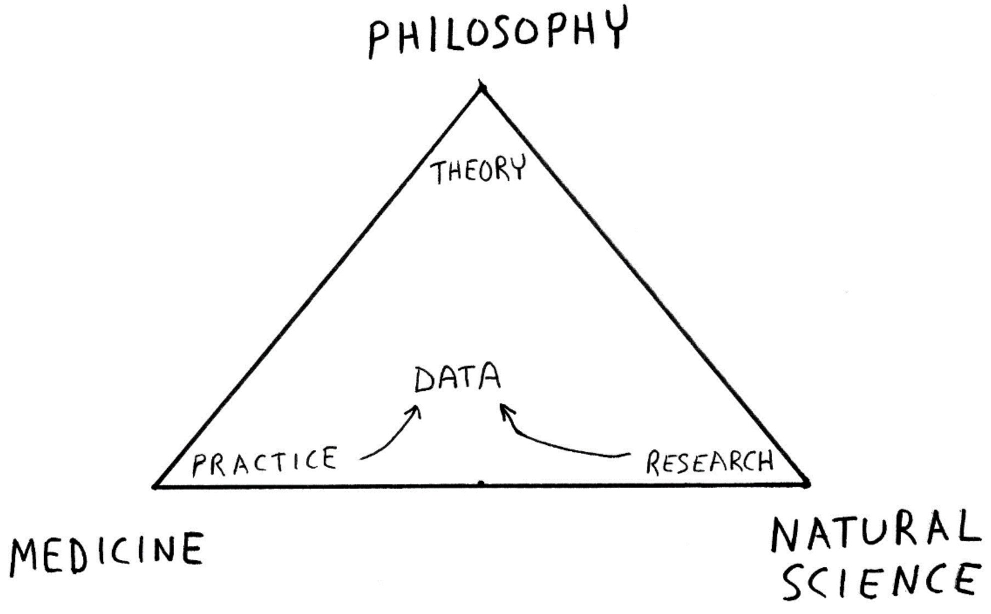


Data and/or Theory?

Research is the way that we discover what is true, and clarify just what it means for us. This is how research produces knowledge. Data constitute one essential component of scientific knowledge. The data come from research (and clinical practice). The other component of scientific knowledge is *theory* (a coherent description and explanation). A theory is a statement of a plausible relationship between variables (e.g., one causes another, one predicts another). Theories help us understand the data, predict future data, and/or control behavior.

Visualize the triangular field of psychology as a pyramid of knowledge. The base of the pyramid must be larger than the capstone. The base of the pyramid is formed by the data, while the capstone is theory.

Psychology's Pyramid of Knowledge



Data without theory are meaningless trivia. Theory without data is idle speculation. Here is the role of theory within psychology: to answer these questions.

- Why do people do what they do? (understanding)
- Can we predict what people will do next? (prediction)
- How can we influence people's thoughts and actions? (control)

This video goes into greater depth describing the role of theory in science.

Hypotheses

Most psychological research begins with a specific question. Usually that question is based upon some theoretical or assumed relationship between variables, and generates a specific prediction. A *hypothesis* (plural hypotheses) is what we call a specific prediction that guides research. Hypotheses are specific predictions about the data we expect to obtain. Don't call them "educated guesses."

Theories generate hypotheses: if this theory holds, then we should expect these data to emerge from this research. The hypothesis is to research what diagnosis is to clinical work: a starting point for the treatment proposed at present, based upon a past fund of knowledge, and confirmed (or not confirmed) by future results. If the data do not fit the hypothesis, then we must rethink the theory that generated that hypothesis. This is Karl Popper's principle of falsifiability: scientific theories must be capable of generating empirically testable hypotheses. A hypothesis that could not conceivably be tested by empirical data would not be scientific. (Claims about religious doctrines would be beyond the realm of scientific investigation.)

Pseudoscience

Pseudoscience (phony science) has claims that resist empirical testing of specific hypotheses. There are many claims about human nature that most psychologists regard as unscientific.

Pseudoscience	Claims
Astrology	Time of birth predicts a person's traits; position of planets determines one's daily fortunes.
Lunar effect	A full moon increases aberrant behavior
Graphology	Handwriting allows prediction of people's future success in relationships and occupations.
Dianetics	Painful memories create "engrams" which can be eliminated by "auditing" while holding electrodes
Conversion (reparative) therapy	Homosexuality is a mental illness that can be treated with psychotherapy.
Parapsychology	Extra-sensory perception, channeling with disembodied spirits, psychokinesis
Biorhythms	There are physical, mental and emotional cycles that can predict human performance.
Homeopathy	Use extreme dilutions of that which brings about the symptoms in order to cure the disease.

However, if any of the above claims were to attain sufficient confirmatory evidence from properly designed surveys and experiments, then the status of these claims would be re-evaluated, and regarded as scientific.

One factor that keeps a pseudoscience going is that it yields the predicted results some of the time, and those will be the times that are most remembered.

"It is the peculiar and perpetual error of the human understanding to be more moved and excited by affirmatives than negatives."

-- Francis Bacon

(When Bacon speaks about the “negatives” he is not referring to bad outcomes, but to outcomes that do not support the theory. Scientists and medical professionals use “positive” to mean present, and “negative” to mean absent.)

Also keep in mind that the major source of evidence cited by pseudoscientists does not come from carefully controlled experiments or comprehensive statistical surveys, but from nice little stories about Joan S. from Atlanta and Peter K. from Brooklyn and Mary Z. from Omaha. These are not even clinical case studies but *anecdotes* selected for their exceptional rather than typical outcomes.

"The plural of anecdote is not data. An anecdote is something that once happened to you, or to your uncle, or to your uncle's accountant. It is too often an outlier, the memorable exception that gets trotted out in an attempt to disprove a larger truth."

-- Levitt & Dubner (2014)

Another technique of pseudoscience is the use of analogy rather than scientific theory in explaining causal relationships. Analogies should not take the place of empirical data. Analogies should only serve to help us comprehend a theory. Analogies provide a creative perspective for viewing complex relationships, but they do not substitute for empirical evidence supporting that such a relationship does, in fact, exist.

In most cases when pseudoscientific claims have been subjected to empirical data (carefully gathered, in great quantity, and cautiously analyzed) the claims just don’t stand up. We remember perhaps a dozen amazing claims of psychics that did come to pass, but forget the thousands of predictions that did not come about.

	Real science	Pseudoscience
Data must be	Empirical	Interesting
Proof cited	Experiments, Surveys	Anecdotes
Claims are	Cautious	Exaggerated
Replication	Encouraged	Ignored
Authors tend to be	University & hospital affiliated	“lone wolf” researchers
Uses explanations from	Established theory	non-mainstream theories, tradition, “common sense”
When hypotheses are not confirmed	Theories are challenged	Data are re-explained by ad hoc explanations
Where to find reports on of this research	Peer-reviewed journals and conferences	Infomercials websites, social media, advertisements
Authority relies on	Quality of research conducted	Charismatic leaders
Distinction between observation and inference	Clear	Obscure

Many people think of psychology, and other social sciences (e.g., sociology, economics, political science) as pseudosciences, or perhaps not much better than common sense. Compared to the "hard" sciences of physics and chemistry, psychology may appear to be less precise, mostly due to the fact that protons are easier to predict and control than human beings. Psychology is a real science only to the extent that it follows the scientific method, basing its claims about human nature on data from experiments and surveys, and theories that have survived hypothesis testing.

"Common Sense"?

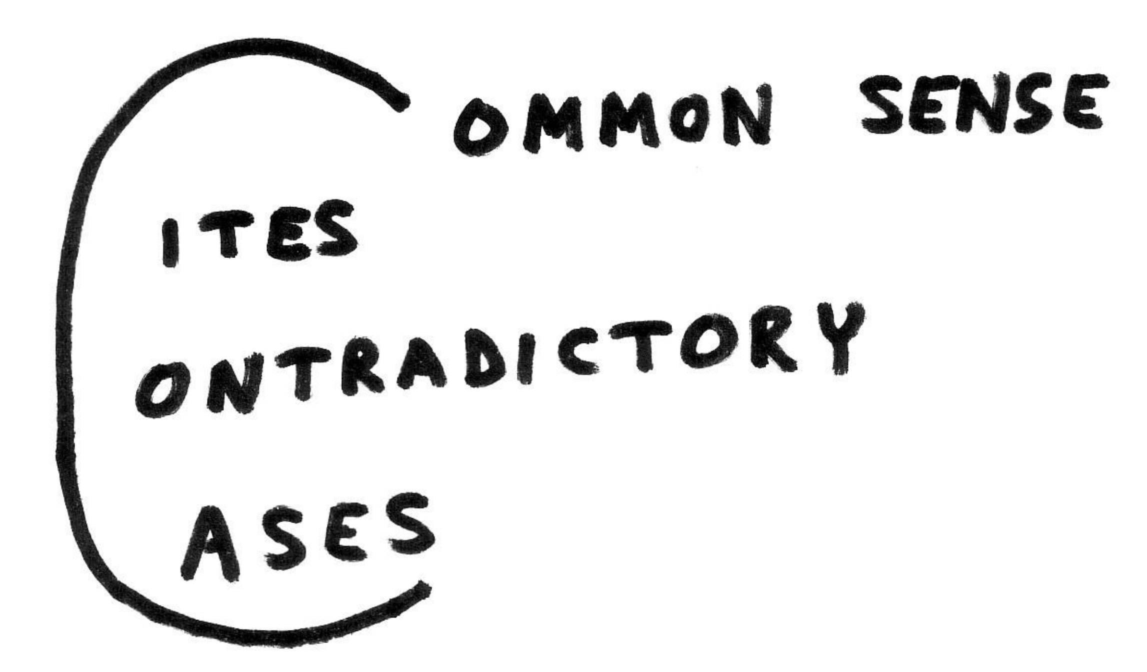
Similar to pseudoscientific claims are those fixed ideas that many non-scientists have about what causes what. These common sense notions can lead to real science if they generate specific hypotheses that survive empirical testing. Unfortunately, many common sense ideas are too vague to formulate empirically testable hypotheses.

Common sense notions about human behavior survive because of a reliance on *introspection* (self-reflection on our own thoughts, feelings, and behaviors) as well as *anecdotes* (especially those confirmatory examples that come to mind). If the data seem to contradict the theory, the common sense theory employs *ad hoc* explanations (which are loose enough to explain any possible result). Using an ad hoc explanation means that the original theory just gets more convoluted rather than getting rejected.

Of course, when common sense (or even pseudoscientific) theories can generate hypotheses that survive empirical data, those theories gain scientific status. However, because most common sense notions have not been established on a scientific basis, it is best not to use the term *common sense* in this course, and when you see it written, regard it as meaning "what is widely assumed by non-scientists."

The proper approach for scientists is to regard pseudoscientific claims and common sense claims skeptically, as not yet verified. Similarly, scientists must regard all scientific theories as merely tentative, vulnerable to future data which may require those theories to be revised or rejected.

Common sense is not science



Singular noun	*Plural noun*	*Adjective*
Analysis	Analyses	Analytic
Bias	Biases	Biased
Crisis	Crises	Critical
Criterion	Criteria	
Datum	Data	
Diagnosis	Diagnoses	Diagnostic
Hypothesis	Hypotheses	Hypothetical
Medium	Media	
Neurosis	Neuroses	Neurotic
Phenomenon	Phenomena	
Prognosis	Prognoses	Prognostic
Psychosis	Psychoses	Psychotic
Stimulus	Stimuli	

Constants & Variables & Operational Definitions

A *constant* is a measure that does not change within a sample (or we could describe a constant for a population or group). A *variable* is a measure changing as we move from measuring one person to another. For example, if all members of our sample were male, gender would be a constant. If both males and females are in the sample, then that is a variable that we should measure.

There must be at least one variable (usually, a dependent variable) that each subject within the sample is measured on. (If there are only constants, we cannot test any hypotheses). How that variable is actually measured (e.g., categories, levels, ranks, numbers) constitutes the operational definition of that variable.

Inference

This word comes from the verb *to infer.* Inference is the process of reasoning from something directly observed to something else not directly observed. Psychologists observe behavior and then make inferences about why the person (or animal) behaved in that way. Emotions, motives, and abilities are never directly observed, but only inferred. Here are some examples of inferences that psychologists or you yourself might make.

OBSERVATION	INFERENCE
The patient scored high on the depression scale.	The patient is feeling very depressed.
The cat went to the water bowl before going to the food bowl.	The cat is more thirsty than hungry right now.
That guy plays his music too loud.	He is a jerk.

Many times, we don't directly measure the variables we talked about. Rather, we measure a specific, observable behavior (or presumed outcome of behavior) and then attempt to infer the level of the variable itself. Sometimes the variable (e.g., a personality trait) is a mere theoretical *construct.* This raises questions about the validity of measurement based upon inference. Are we really measuring the construct we claim to measure, or some other variable that was just easier to measure?

This video goes into greater depth about the role of inference in science.

Causation

Science tries to explain the natural world with theories of cause and effect. Sometimes we observe an effect, and infer a likely cause.

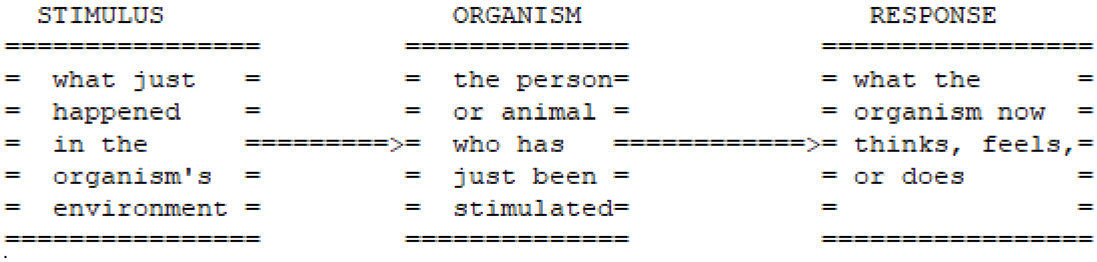
OBSERVATION *effect*	INFERENCE *cause*
The little girl is crying.	She probably fell and got hurt.
That worker is behaving in an unsafe manner.	He has not received sufficient training.
Customer purchases of the new product have increased.	The ad campaign must be effective.
The patient is still depressed.	The dosage is too low.

Of course, if the cause was not essential to produce the effect, we could be mistaken, for there may be some other cause of the observed behavior. Perhaps the little girl was not able to use the swing because another child cut in front of her: she was not physically hurt, but her sadness was due to disappointment.

Sometimes we observe a cause, and infer a subsequent effect.

OBSERVATION *cause*	INFERENCE *effect*
That little boy is being badly beaten by his father.	He will grow up to become a serial killer.

Of course, if the cause is not always adequate to produce the effect, these predictions can be mistaken. Predictions are much easier in a science like physics, where all hydrogen atoms always react in the same way. In psychology, we must keep in mind that people do not merely react, but they respond. Between the cause (an environmental *stimulus*) and the effect (the *response*) is an *organism* (the *subject* of our research, a person or an animal). The stimulus is always something external, a change in energy that the organism can perceive (e.g., a loud sound). The stimulus is not an internal drive (e.g., hunger). The organism is a person or animal perceiving the stimulus who then creates a response. The response is what the organism does (e.g., action, speech, scores on a test). The stimulus *elicits* a response; the organism *emits* a response.



```
  STIMULUS                  ORGANISM                  RESPONSE
================         ==============            =================
=  what just   =         =  the person=            = what the      =
=  happened    =         =  or animal =            = organism now  =
=  in the      =========>=  who has   ============>= thinks, feels,=
=  organism's  =         =  just been =            = or does       =
=  environment =         =  stimulated=            =               =
================         ==============            =================
```

Dependent Variables

In analyzing causation, there are two types of variables: dependent (observed effects) and independent (usually understood to be the possible causes of those effects). In psychology, the *dependent* variable will always be some form of behavior. *Performance* is just a measurement of behavior according to some standard (e.g., speed of running a course, units produced on an assembly line, sales made). Going back to the stimulus and response model, it can be said that in psychology, the dependent variable corresponds to the response. Here are some examples of dependent variables and their operational definitions. This video discusses variables in greater depth.

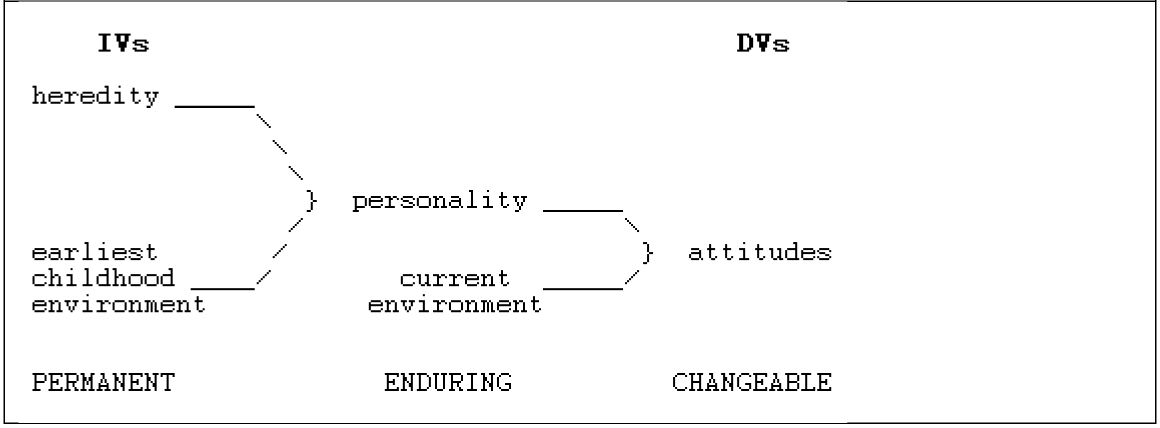
ORGANISM	DEPENDENT VARIABLE	OPERATIONAL DEFINITION
Rat	Performance running a maze	Number of seconds it took to get through the maze
Voter	Attitude about a political candidate	Whom the voter says that she will vote for
Consumer	Decision to purchase a product	Whether or not the consumer purchases the product
Worker	Absenteeism	How many times last year the worker did not show up for a scheduled shift
Patient	Depression	Score on a valid and reliable depression scale

All decisions made by the subjects are dependent variables, but not all dependent variables are decisions. Some outcomes (effects) are not intended by the subjects. One example of this would be mortality or a persisting mental illness. These should be regarded as dependent variables in that they are the results of (ineffective or dangerous) treatment, but they are not what the subject preferred.

Attitudes

Learned habits for responding to social stimuli are known as *attitudes*. These must not be confused with personality traits. Attitudes are not as permanent or as internally consistent as personality traits. Traits are supposed to be characteristic of the individual, regardless of the situation. Attitudes are more influenced by the situation. The relationship between attitudes and personality traits is seen below.

Relationship between PERSONALITY and ATTITUDES



The common use of the term "attitude" often fails to appreciate this distinction. Whenever you hear someone say "He has a bad attitude" (if it is about everything, it is not an attitude contingent upon the object, it is an enduring trait of the subject). Remember: personality is composed of traits that are enduring and integrated. Attitudes are social habits that are both diverse and changeable.

As this video shows, an attitude is always about a specific thing, an object, and describes the subject's understanding of that object, the emotional evaluation of that it, and the subject’s predisposition to act in a certain way toward it. So, each attitude can be dissected into three components: cognitive, affective, and behavioral.

In psychology, whenever we are talking about actions, decisions, choices, attitudes, performance, or scores on tests, we are talking about dependent variables. Of course, each of these variables could also be seen as a cause of some other event further down the chain. The rat may receive a reward for running the maze quickly, the worker might get fired for his absenteeism. However, in psychology, the dependent variables are behaviors, not the later consequences of the behaviors.

Independent Variables

These are the potential influences upon behavior. Some independent variables are stimuli coming in from the environment. Here are some examples of such independent variables.

ORGANISM	INDEPENDENT VARIABLE *(stimulus factor)*	DEPENDENT VARIABLE *(result influenced by the independent variable)*
Rat	Shape of the maze	Performance running the maze
Voter	Campaign materials	Attitude about political candidate
Consumer	Advertisement	Decision to purchase a product
Worker	The surf report	Absenteeism
Patient	Death of his wife three weeks ago	Depression

All stimuli are independent variables, but not all independent variables are stimuli. Another type of independent variable would be something in the organism's background that also influences the organism's behavior. This can include hereditary factors or experiences during a formative time in the organism's life, such as early childhood. This video goes into greater depth about variables.

ORGANISM	INDEPENDENT VARIABLE *(background factor)*	DEPENDENT VARIABLE *(result influenced by the independent variable)*
Rat	Age of rat	Performance running a maze
Voter	Raised by parents who were strict Republicans	Attitude about a political candidate
Consumer	Female gender	Decision to purchase a product
Worker	His father was an alcoholic	Absenteeism
Patient	He was orphaned in childhood	Depression

The goal of science is to understand, predict, and control. Science tries to explain things in terms of cause and effect. Psychology is the science of behavior, so it tries to explain behavior in terms of independent and dependent variables.

Prediction

Not all psychological research is able to identify the relationship between variables such that one is clearly indicated as the cause (independent) of the other (dependent). Sometimes, the best we can do is to determine that the variables are associated (correlated). This at least allows us to predict something about the level of one variable from knowledge of the other.

The variable we are trying to predict is known as the *criterion variable*. This must be a dependent variable. All criterion variables are dependent, but not all dependent variables are used as criterion variables. Usually, the criterion variable is some future behavior or outcome that we are trying to predict from knowledge of other (present or past) variables.

The variables we use to try to make that prediction are known as the *predictor variables*. These predictor variables can be (past or present) independent variables (e.g., organismic background factors, stimuli to which the organism was exposed) or even dependent variables (e.g., past or present measures of the organism's performance). Indeed, one rule-of-thumb in industrial psychology is that past behavior is usually the best predictor of future behavior. One rule-of-thumb in political psychology and consumer psychology is that past decisions are usually the best predictor of future decisions.

ORGANISM	PREDICTOR VARIABLE	CRITERION VARIABLE *(outcome)*
Rat	The rat ran the maze quickly yesterday. DV: previous outcome	The rat will probably run the maze quickly again today
Voter	She voted Republican last time. DV: previous decision	She will probably vote Republican again this year
Consumer	She is a woman. IV: background	She will probably look for clothes in the women's section
Worker	His supervisor has evaluated him as "irresponsible." DV: assessment	He will probably be absent more often than the other workers
Patient	He is now getting psychotherapy. IV: treatment	Depression will probably subside in eight weeks

Null Hypothesis

So, just how do we know, as scientists, when we have sufficiently “proved” our hypotheses? Here are some claims that could be tested with empirical data.

- This sample has a high level of depression
- Losing one’s job causes depression in middle-aged men
- Depressed people are pessimistic about the future
- This new test is a valid measure of depression
- This new medication is an effective treatment of depression

One technique for achieving a certain level of confidence in our proof is usually accomplished through a process known as *null hypothesis* significance testing. It impresses most people as a backwards way to do things. We really don’t prove that our hypothesis is true, or even likely to be true: we try to show that another explanation is unlikely to be true. The null hypothesis is the opposite of what we are trying to prove. We are trying to prove something like

- This sample has such a high level of depression that it could not be explained by random variation within the normal population.
- The difference in depression scores between a group of men who just lost their jobs, and those who have not just lost their jobs, is greater than that which pure chance might explain.

- The correlation between depression and pessimism is more of a trend than random variation could explain.

- The correlation between this new depression scale and a previously established measure of depression is more of a trend than random variation could explain.

- The difference in depression scores between a group of men who were treated for four weeks with anti-depressant medication, and a group of men who were just given a placebo, is greater than that which pure chance might explain.

In each of the above examples, the null hypothesis would state that there is no real difference or trend that could not be explained by random variation (pure chance, luck). In that sense, the null hypothesis says that we proved nothing. So, in order to prove something, the first step is to show that the null hypothesis is unlikely, and reject it. This video explains the logic of null hypothesis testing.

We use *inferential statistics* to calculate (or estimate) the probability of the null hypothesis being able to explain the observed data. Probabilities are represented by decimal numbers that range from 0.00 (which stands for something completely impossible) to 1.00 (which stands for something completely certain).

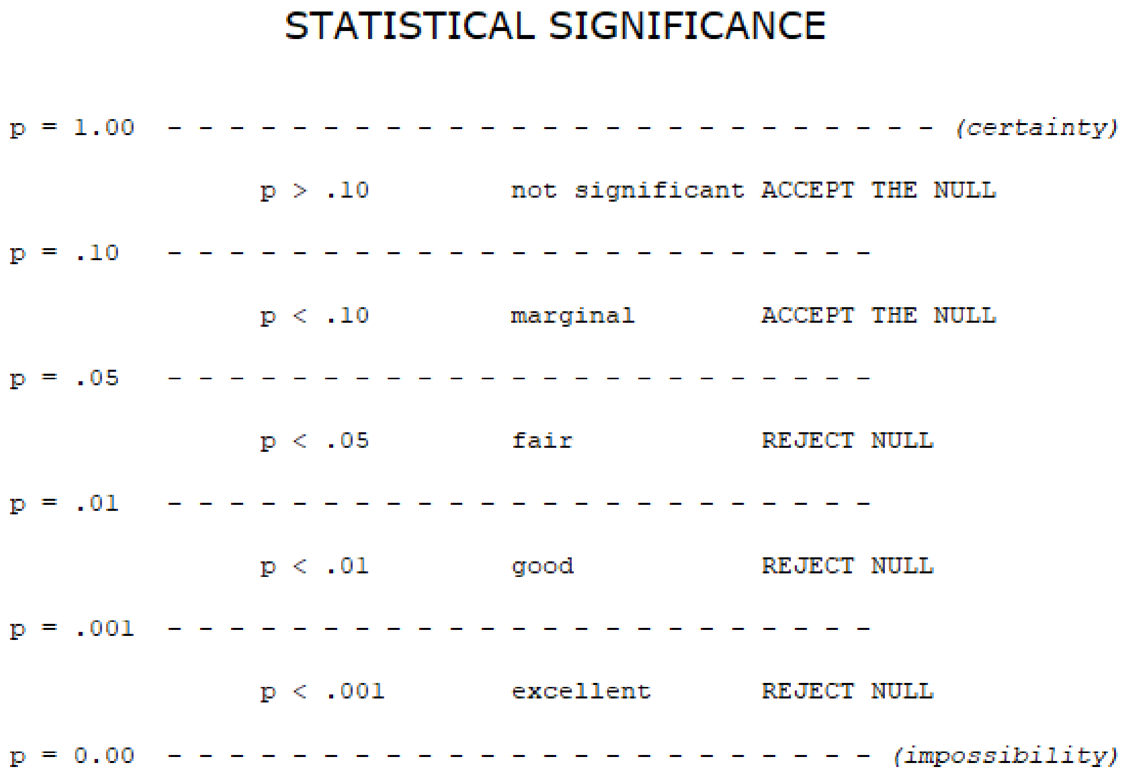
Statistical Significance

For example, suppose I make some pseudoscientific claim (e.g., psychokinesis) such that if you flipped a coin, I could use my mind power to make it come up heads. If you are thinking like a scientist, you would start off being skeptical of my claim, and demand an empirical demonstration. So, you flip a coin, and I say the magic word, and lo and behold, the coin comes up heads. I say "See, I told you so." But then you say, "That was just pure luck because you had a fifty-fifty chance of getting it right." What you have just done is accepted the null hypothesis as your explanation. You looked at the fact that the probability is 50% ($p = .50$) and you concluded that my data were not statistically significant. So, we flip the coin again, and again it comes up heads, but still you are not convinced, because ($p = .25$), there is a one in four chance that I could get two in a row. So you stick with the null hypothesis, claiming that I am just lucky. Even after three flips resulting in three heads, most students would still say "still probably just luck" (because $p = .125$ at this point).

However, there would come a point at which you would say "No one is that lucky, there is something else going on here." At that point you have rejected the null hypothesis as an explanation, because its probability was too low. After you have rejected the null, then some other explanation (e.g., fraud, psychokinetic ability) must be considered. To use a legal analogy, just as we must presume innocence until we prove guilt beyond a reasonable doubt, so we can only reject the null if the probability for the null becomes unreasonable. Until we get to that point, we should doubt the proof for claims of a relationship between variables.

We need to come to some agreement as to where to draw the line, when to reject the null hypothesis as an explanation for our data. Most editors of psychology journals have regarded $p < .05$ as the cutoff. If the probability of the null is greater than .05 ($p > .05$) then we do not reject the null and we must admit that the data are not statistically significant. If the probability is less than .05 ($p < .05$) then we reject the null with fair confidence that we have really proved something. The smaller the probability of the null hypothesis, the more confident we can be about the significance of our findings.

Null hypothesis testing has its limitations (and its critics within science). Of course, it is possible that I could be very lucky and call five or even ten flips of a coin by pure chance. The important thing to remember is that statistical significance (even at an excellent level of p < .001) does not prove a causal hypothesis, it merely shows that another competing explanation (pure chance) is unlikely. (This video explains the following significance table in greater depth.)



STATISTICAL SIGNIFICANCE

```
p = 1.00  - - - - - - - - - - - - - - - - - - - - - - - - -  (certainty)

                p > .10        not significant ACCEPT THE NULL

p = .10   - - - - - - - - - - - - - - - - - - - - - - -

                p < .10        marginal        ACCEPT THE NULL

p = .05   - - - - - - - - - - - - - - - - - - - - - - -

                p < .05        fair            REJECT NULL

p = .01   - - - - - - - - - - - - - - - - - - - - - - -

                p < .01        good            REJECT NULL

p = .001  - - - - - - - - - - - - - - - - - - - - - - -

                p < .001       excellent       REJECT NULL

p = 0.00  - - - - - - - - - - - - - - - - - - - - - - -  (impossibility)
```

Many statisticians caution that we should not claim that a high p value (e.g., p > .50) proves no relationship between the variables. It would be better to state that we failed to prove any relationship between the variables. This is because several things (such as sample size) other than the strength of the correlation between the variables can influence p values.

Many statisticians don't even like to use the term "accept the null" preferring instead to say "cannot reject the null." We should not have a problem as long as we remember that accepting (or failing to reject) the null simply means that we admit that we have proved nothing. If by "accepting the null" we mean that we have proved that there is no difference, then that requires Bayesian techniques.

When we speak about statistical significance we should use terms such as *excellent, good, fair, marginal,* and *not significant*. It is confusing to use terms such as *strong* and *weak,* or *low* and *high* when discussing p values (because lower is better), so reserve those terms for describing correlation coefficients.

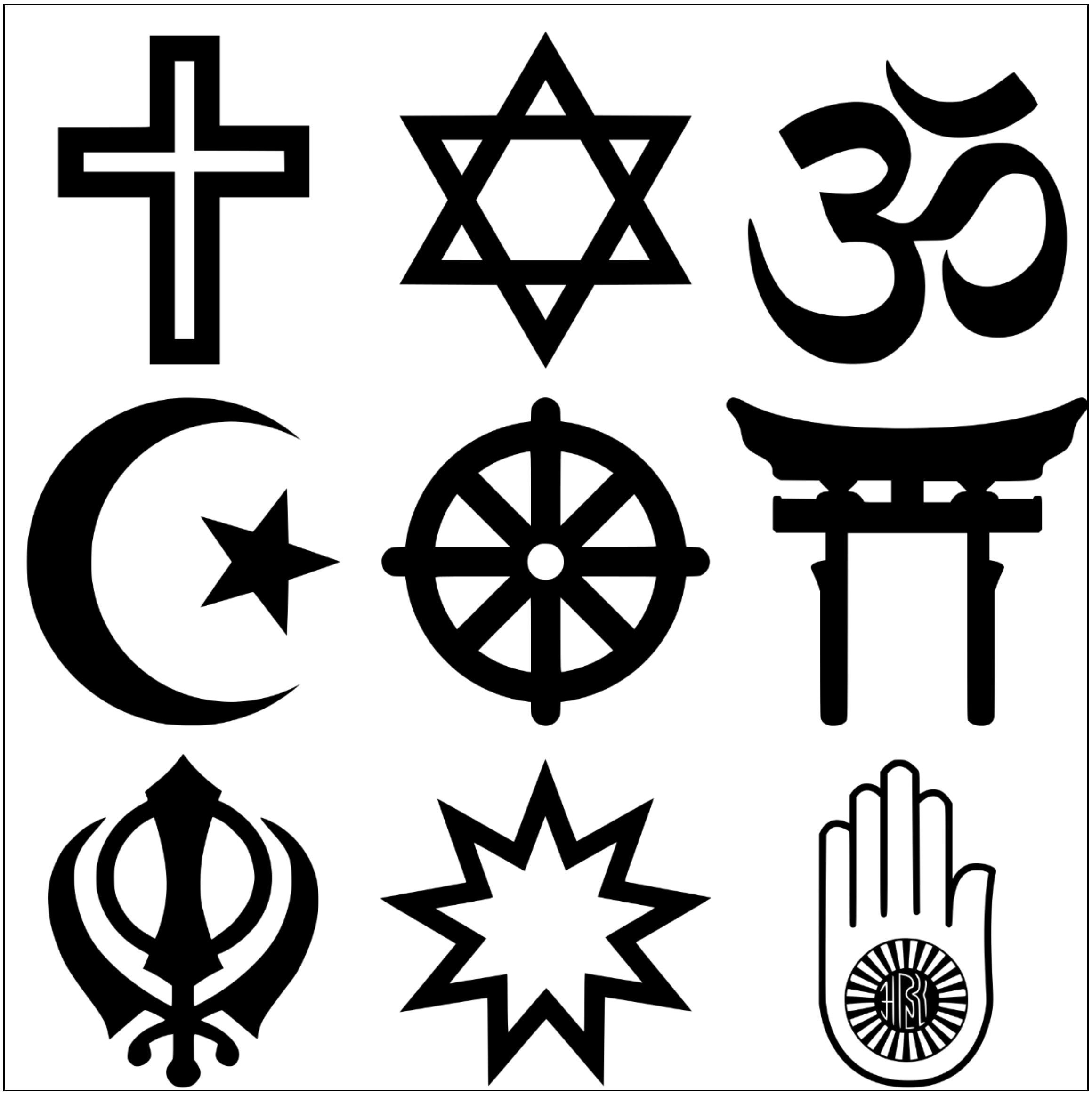
In practice, statistical significance is influenced by several factors. The first is sample size. In general, the larger the sample is, the more significant the data are. If the sample size is small (say, under twenty) it is pretty hard to reject the null. Professional polling organizations, such as Gallup, usually go with a sample size of over a thousand, because that means that a difference of just a few percentage points in the polls will be statistically significant ($p < .05$). Another factor influencing significance is the magnitude of the difference between the groups (or the strength of the correlation coefficient). The stronger the correlation (or the greater the difference between two groups), the more likely it is to be a significant one. It takes a large sample for a small difference to be significant, and it takes a large difference for a small sample to yield significant data. Yet another factor is how much dispersion there is on the dependent variable. High standard deviations may make it harder to achieve statistical significance.

influence on significance	Better	Worse
Difference between groups	Bigger	Smaller
Difference within groups	Smaller	Bigger
Sample size	Bigger	Smaller

Religion

Religion is defined as a system of doctrines, ethics, rituals, myths, and symbols for the expression of ultimate relevance. Doctrines are statements about deities and afterlife (things that we cannot observe with scientific instrumentation). Ethics are guidelines for behavior: what is right and wrong (and such value judgments cannot be verified by the empirical method). Myths are stories about the past, which may or may not be historically true. History is a social science, so the facts behind historical claims must be verified, but myths are retold because of the values they portray. Symbols are emblematic expressions of doctrines, ethics, or myths, and have no real operational definition. Rituals are ceremonies that use symbols to re-enact myths.

This video gives a mnemonic for remembering the definition of religion. This video shows how scientific data and theory can be used to study some aspects of religion.



Some (but not most) scientists are atheists who view religion as not much more than superstition or pseudoscience. Sigmund Freud (the Psychoanalyst) and B.F. Skinner (the Behaviorist) thought that as science came to better understand human behavior, there would be less reliance upon religion.

Conversely, some religious extremists may oppose science. Cult leaders may claim to be the only authority on everything and forbid their followers from consulting science on topics such as evolution, the earth revolving around the sun, or receiving medical treatment. Some traditional religious fundamentalists take scripture (e.g., the Bible, Torah, Qur’an) literally, and contend that scripture contains all that we need to know about human nature, and therefore, we do not need a science of behavior.

On the relationship of religion and psychology, this book takes the middle position: there is no contradiction between the two because they employ different methodologies in coming to conclusions about human nature. Psychology and other sciences use the empirical method of observation. Religion gets its knowledge from revelation: scripture, a prophet, a pope, etc. Science tells us what people *are* like, while religion tells us what people *should* be like. Psychology searches for techniques to promote mental health, while religion seeks salvation. It is the contention of this book that one can be a devout Christian, Jew, Hindu, Jain, Sikh, Shinto, Confucian, Daoist, Zoroastrian, Muslim, Wiccan or Buddhist, and also be a good scientist.

The religiously devout should not be concerned that psychology, or any other science, is going to conclude that God does not exist, or come up with another formula for saving one's soul. Remember, the definition of religion as a system of doctrines, ethics, rituals, myths and symbols for the expression of ultimate relevance. Science cannot answer (one way or the other) any of the following questions.

- Which deities (if any) merit worship?
- Is abortion morally wrong?
- Should baptism be performed using full immersion?
- Should infants be baptized?

- What is the deeper meaning of the story of Noah?

- Should we pray in front of statues of saints?

- Does being a Catholic give you a better chance of getting into heaven?

- At death, does the soul go to heaven, hell, purgatory, limbo, get reincarnated, or merely sleep in the grave until the resurrection?

There is no way we can set up an experiment or a survey so that it will provide an empirical test of a hypothesis about any of these components of religion. If we have to accept a null hypothesis, that does not mean that God does not exist. Doctrinal statements about deities are usually *ad hoc* hypotheses that can explain whatever empirical data might be encountered. If we pray to God for a miracle, and nothing happens, that does not prove atheism. "Perhaps God knows that it is better for us to endure a challenge, and has therefore decided not to perform a miracle."

So, scientific statements cannot prove or disprove religion, and religious statements should not be regarded as science (at least, not until such a statement has been confirmed by empirical means, not just by revelation).

	SCIENCE	RELIGION
Method	Empirical observation	Revelation
Reality is an interaction of	Natural phenomena	Spiritual beings
Truth as	Valid data	Enduring values
Human nature	The way it is	The way it should be
View of the past	History	Myth
Relevance of human action	Technology	Ritual
Focus on	Variables to be measured	Symbols to be revered

Religion only oversteps its bounds and wanders into the territory of science when religion starts making empirical claims (e.g., that life on earth is only six thousand years old). As long as religion talks about the relevance of values and the characteristics of spirit beings who have no coordinates in space and time, then science cannot perform any measurements to challenge religious statements.

While science must remain agnostic, insofar as it cannot pass judgment on religious questions, individual scientists do not have to be agnostics. Most psychologists, psychiatrists, and psychotherapists are not atheists, but have some religious affiliation. Indeed, many Catholic priests, Protestant ministers, and Jewish rabbis blend modern psychotherapeutic techniques with traditional spiritual counseling in what is known as pastoral care.

While psychology and other social sciences can pass no judgment on the truth claims of religious doctrine, ethics, rituals, myths or symbols, these social sciences can study individuals’ religious attitudes and behaviors as dependent variables, and note the relationship with various predictors (e.g., ethnicity, socio-economic status) or criterion variables such as life satisfaction or political attitudes. The following hypotheses can be empirically tested…

- The more religious that people are, the greater their life satisfaction.

- People born after 1980 report lower levels of religiosity than do persons born before 1945.

- Married couples sharing the same religious affiliation are more likely to remain married than couples who are from different religious affiliations.

- Hindus in the U.S. have higher educational attainments than do Jehovah's Witnesses.

- People who describe themselves as Evangelical Christians are more likely than atheists to oppose same-sex marriage.

- People who have the personality trait of openness to new experiences are more likely to undergo religious conversion.

Notice that most of these hypotheses are more along the lines of prediction rather than causation. Even if we assume that the above stated associations hold, it is not clear that

- If you become more religious, your life satisfaction will go up.

- The level of religiosity of today's Millennials is due to being born in the 1980s and 1990s, rather than their current age.

- Religion (rather than cultural similarity) is what is keeping married couples together.

- If you convert to Hinduism, your income will go up and if you become a Jehovah’s Witness, your income will go down.

- Religious affiliation causes political affiliation (rather than the other way around).

- Personality caused the conversion (rather than was the result of the conversion).

Being a good scientist is, first and foremost, realizing the limitations of science. Know that even when you can reject the null, it does not always allow you to infer cause and effect. The rest of this book gives guidelines on how to do that.

Chapter #2: Literature Review

To find out what has already been written on your topic, you should conduct what is known as a literature review. This is also known as *secondary research* (with *primary research* being the raw data that you collect in your own qualitative and quantitative investigations).

It is alright to begin secondary research with *Wikipedia*, because it is usually clear and well-organized, but realize that its content may change rapidly, and may contain errors of fact (especially on controversial topics). Wikipedia, like popular websites, magazine articles, and self-help books, lacks a formal system of *peer review*, where a panel of experts in an area of knowledge approve the content before it is published. Peer review is the main characteristic of scholarly journal articles, but can also be found in most conference presentations at professional societies (e.g., American Psychological Association, Association for Psychological Science), books published by academic publishers (e.g., university presses), and even the better encyclopedias (e.g., *Britannica*).

Start here, but don’t stop here

Go to the Library

You need to go to the library. Today’s college students think of the library as the place where you go to collaborate with members on your project team, or study together, or relax on a soft chair while you check your incoming texts. Before libraries became the place to do those things, libraries were simply repositories for books, especially reference books that were too rare or too expensive for individuals to purchase. Much can be found in Wikipedia or on other internet sites, but there are still valuable reference books that you need for your literature review. You need to physically go to the library and look at the reference books there.

Here are links to the book catalogs of local libraries, such as the catalogue at the Crafton Hills College library. This catalog is for both colleges in the San Bernardino

Community College District: Crafton Hills College and San Bernardino Valley College. The latter has more books to check out. If it is not convenient for you to go all the way over to the SBVC campus, you can have circulating books sent over to the CHC library, and when it is time to return those books, you can just drop them off at the CHC library as well, thus avoiding any trips to the more distant campus.

Even if the location and hours of our own college libraries are not convenient for you, go there at least once and get an ILEAC card. This allows you to use other libraries in the Inland Empire Academic Library Cooperative.

The Webb library at Loma Linda University is the best on medical topics (and some religious topics) within a fifty mile radius. It has the added benefit of being open early Sunday morning (but closed all day Saturday).

The Armacost library at University of Redlands is easy to get into and very laid back and comfortable with places to relax or have coffee. You won't need special ID to get in, but you may need an ILEAC card to check out materials. The staff are usually very helpful.

The Redlands City Smiley Public Library is architecturally impressive and parts look like a museum. There is a good collection of reference books for a relatively small, public library. I prefer it over the city libraries of larger neighboring cities (e.g., San Bernardino, Riverside). You have to go to a large city like Los Angeles or San Diego or Phoenix to get a more comprehensive city library.

The main San Bernardino City Feldheim Library is easy to get to by public transportation, corner of 6th & E streets.

Most of the local libraries in the Inland Empire are part of the San Bernardino or Riverside County system (not to be confused with the city libraries mentioned earlier). The San Bernardino County Library system has many local branches. My own preference (especially for reference works) would be the branches at Highland and Fontana.

You would have to go to the specific branch to look at a reference book, but when it comes to circulating books (i.e., those that you can check out) you can go to any convenient local branch (e.g., Yucaipa, Mentone) and arrange for the book you want to be sent to that local branch, where you can pick it up. When you are done with that book, you won't have to drive to Fontana to return it, but can return it at your local branch.

Here are some of the better encyclopedias available in those aforementioned libraries.

New Encyclopedia Britannica
030 B77e Smiley Public Library

Encyclopedia of Psychology
Edited by Alan Kazdin, American Psychological Association
Ref BF31 .E52 2000 Crafton Hills College library
University of Redlands library

Encyclopedia of Psychology
Edited by Raymond Corsini
BF31 .E52 1994 Crafton Hills College library
University of Redlands library

Encyclopedia of psychology and religion, edited by
David A. Leeming, Kathryn Madden, Stanton Marlan
New York ; London : Springer
BL53 .E43 2010 University of Redlands library

The Encyclopedia of Positive Psychology
edited by Shane J. Lopez.
Chichester, U.K. ; Malden, MA : Wiley-Blackwell
BF204.6 .E53 2009 University of Redlands library

Encyclopedia of Multicultural Psychology
Yo Jackson, editor.
Thousand Oaks, CA. : SAGE Publications
GN502 .E63 2006 University of Redlands library

Magill's Encyclopedia of Social Science: Psychology
Edited by Nancy A. Piotrowski
Pasadena, CA: Salem Press
BF31 .M33 2003 University of Redlands library

The Encyclopedia of Psychiatry, Psychology, and Psychoanalysis, edited by Benjamin B. Wolman, editor,
New York: Henry Holt
RC437 .E49 1996 University of Redlands library

The Gale Encyclopedia of Psychology
Susan Gall, executive editor; Bernard Beins and Alan J. Feldman, Detroit : Gale
BF31 .G35 1996 University of Redlands library

The Blackwell Encyclopedia of Social Psychology
edited by Antony S.R. Manstead and Miles Hewstone
Oxford ; Cambridge, Mass. : Blackwell
HM251 .B476 1995 University of Redlands library

Remember, quality encyclopedia articles are identified by the name of the author, and you should cite them by author's last name.

Search Engines & Databases

In order to find out what has been published in scholarly journals, a convenient place to start would be *Scholar Google* and it really helps to look at the search tips.

To vastly improve your searches with Scholar Google, click on the little triangle just to the left of the spyglass icon to open up advanced search, as this video demonstrates.

One of the largest databases in the social sciences is *SocIndex*. *PubMed* looks at publications in medical journals. The database maintained by the American Psychological Association is *PsycINFO* and covers all of the APA journals. One of the databases maintained by the college is *EBSCO*. Other databases are *JSTOR*, *Proquest*, *Wilson*, *Social Science Research Network*, and *Educational Resources Information Center* (ERIC). If you are a student at Crafton Hills College, you may receive passwords to get into these databases from our librarian.

Another great way to build a bibliography about any topic related to education is to use the National Center for Education Statistics. For help with specific variables and norms, see these.

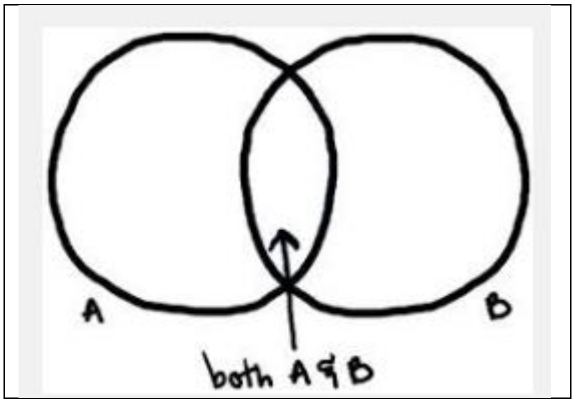
Some journal publishers have come up with easy access to their articles, which can be searched from one site. Here is an example from Taylor & Francis.

Meet with Dr. Brink in his office for help on using the right key words in searching these databases and search engines. Use *Boolean* terms such as OR between the words to expand the search and terms such as AND and NOT to narrow down the search.

Boolean Search Terms	
AND	Returns only results meeting both search terms
OR	Returns results meeting either search term
NOT	Excludes results with this term

	O			E
A	N	D		I
	L			T
	Y			H
				E
			O	R

Think of a Venn Diagram of overlapping circles.



The use of the AND term limits your search to the (small) overlap of terms A and B (both terms), while using the OR term includes the entire area of both circles (either term).

For example, typing in

Watson OR Behaviorism

should return all articles dealing with Behaviorism (even those that talk about Skinner or some other Behaviorist, even if those articles fail to mention Watson) as well as articles authored by some other person named Watson (even if those articles have nothing to do about Behaviorism). Obviously, this strategy of using OR gives you more hits, perhaps too many. If you find that most of these are irrelevant, try one of the strategies mentioned below.

Type in

> Watson AND Behaviorism

which should return only articles by or about John Watson, the founder of Behaviorism. But realize that this AND approach would exclude some otherwise interesting articles about Behaviorism, as well as articles talking about John Watson (if those articles did not mention the specific word "Behaviorism").

Suppose the problem produced by your first search using AND was that there were too many articles about other things names Watson (e.g., the IBM computer that plays Jeopardy and Sherlock Holmes' fictional sidekick). Another solution would be to type in

> NOT Holmes

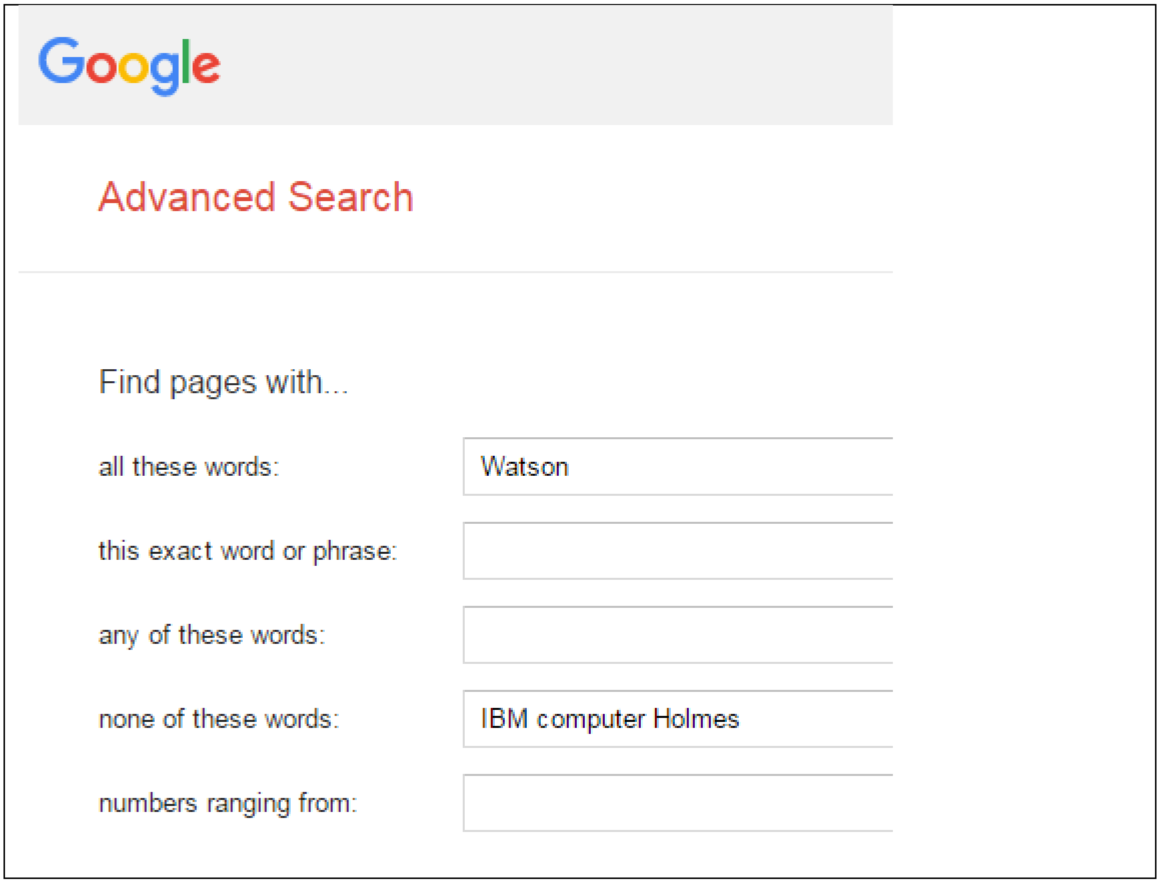
or

> NOT IBM

or

> NOT computer

Some search systems may allow you to type in something like this all at once (as in Google advanced search)



Advanced Search

Find pages with...

all these words: Watson

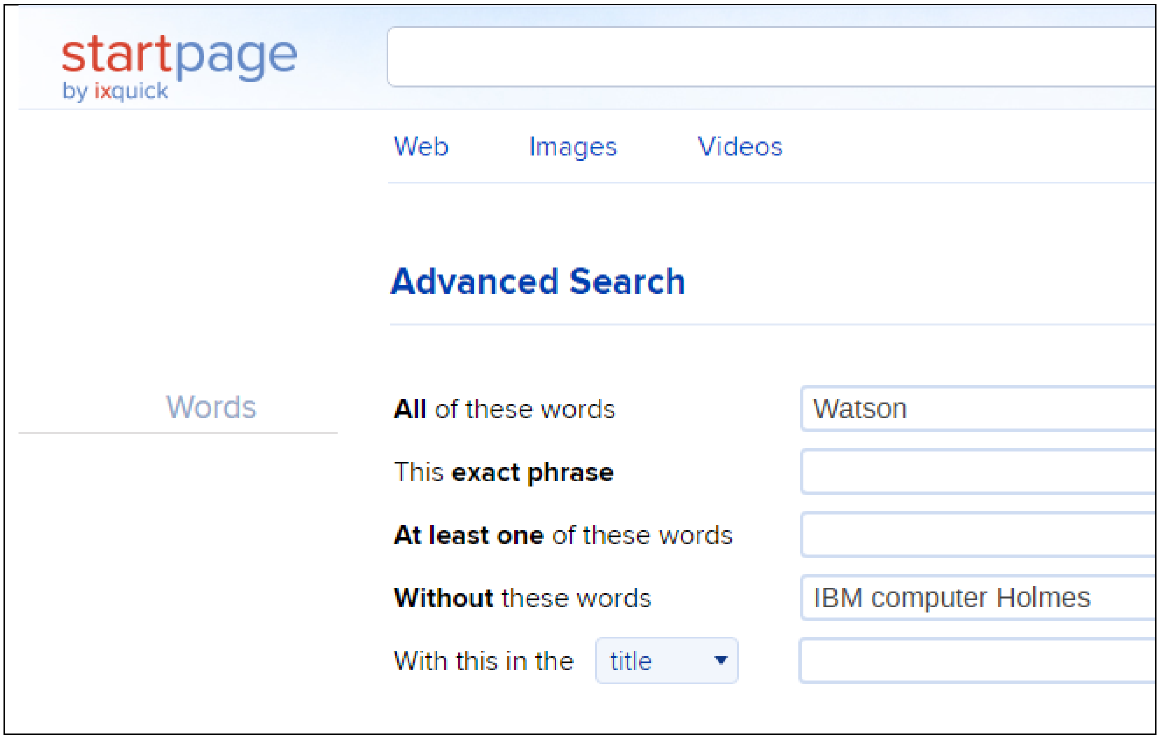
this exact word or phrase:

any of these words:

none of these words: IBM computer Holmes

numbers ranging from:

Or with Ixquick advanced search



All of these words Watson

This **exact phrase**

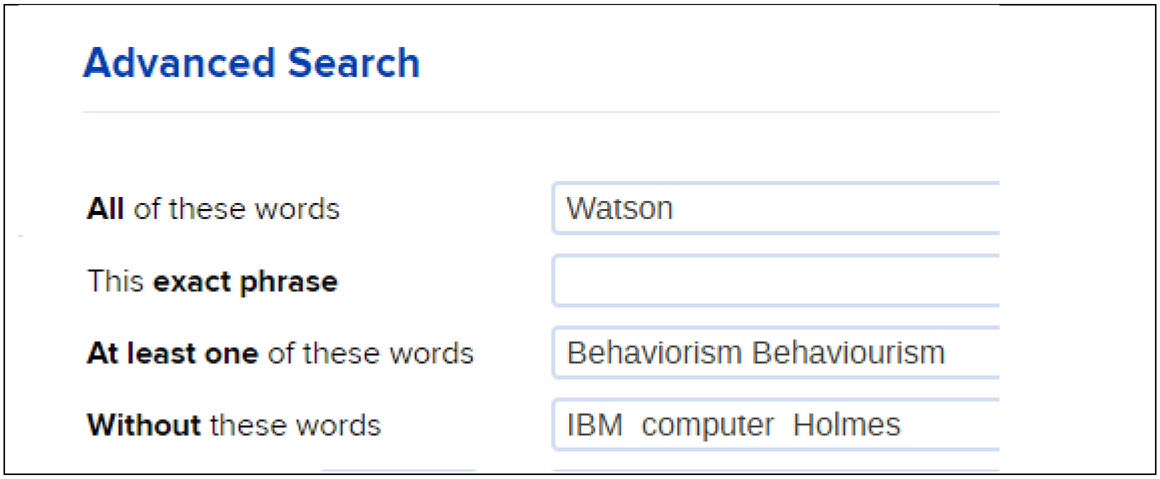
At least one of these words

Without these words IBM computer Holmes

With this in the title

Another approach for narrowing a search is to use quotation marks to return an exact phrase. If we say “John Watson” then we will not return any other Watsons (e.g., Holmes’ or IBM’s). However, some search engines and databases are a little too precise and may not return an article where Watson is only referred to as “John B. Watson” or “John Broadus Watson” or “J. Watson” or “J.B. Watson”.

Also keep in mind that some articles might have misspelled part of the name or used an alternative spelling, such as Behaviourism. (Many older articles used more hyphenated spellings, such as psycho-analysis. Here is a way to cope with this challenge in Ixquick advanced search.



Advanced Search

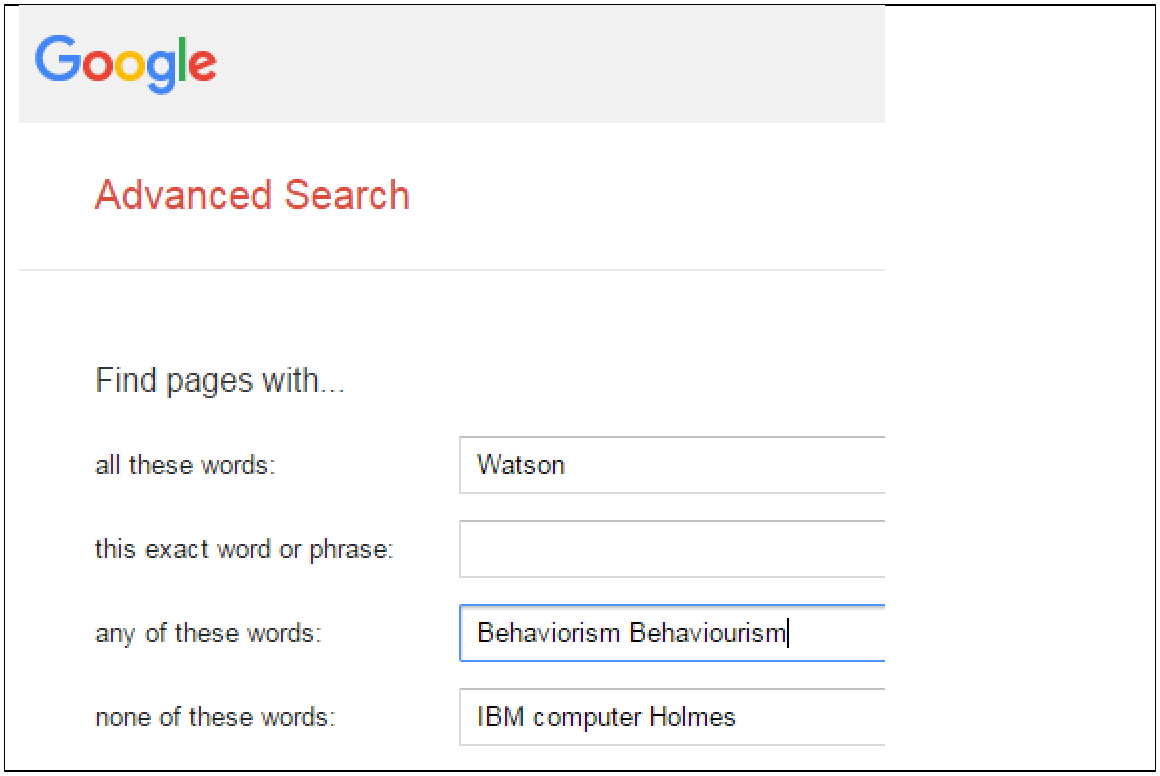
All of these words: Watson

This **exact phrase**:

At least one of these words: Behaviorism Behaviourism

Without these words: IBM computer Holmes

Or Google advanced search



Advanced Search

Find pages with...

all these words: Watson

this exact word or phrase:

any of these words: Behaviorism Behaviourism

none of these words: IBM computer Holmes

Each search engine and database has its own quirks. Play around with them, especially with the wildcard functions. Try typing in

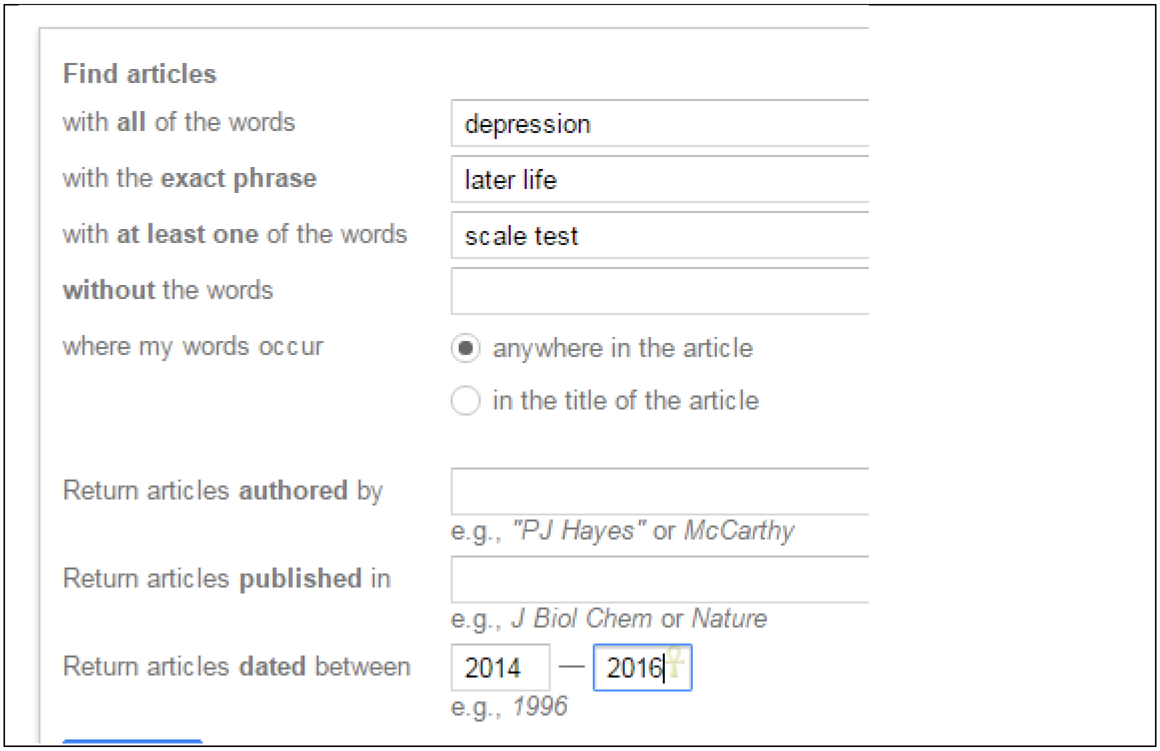
Behavio

Does it return just people and companies with that exact name? How about the website Behav.io? In some search engines and databases, that term would also return Behavior, Behaviorism, Behavioral, Behaviorist, as well as British spellings of these words (which have a U before the R). Some search engines and databases will perform these extra searches with a *wildcard* character after the last letter (usually an *). Typing

Behavior*

into Ixquick gives me no more hits than the single page I got without the asterisk, but with Google, I now get over a half million hits by adding the asterisk.

Another way to limit the large number of results that you see from a data base search is to limit the dates to just the last few years. Especially when dealing with journal articles that you discover on a database, start with the most recent ones first because they will reference other previous important articles that you then look for. For example, suppose you are interested in depression in later life. You want to know what is the best depression scale to use. Look at some of the most recent articles on this topic. Here is how you would use Scholar Google's advanced search to do this.



Find articles

with **all** of the words: depression

with the **exact phrase**: later life

with **at least one** of the words: scale test

without the words:

where my words occur: ● anywhere in the article ○ in the title of the article

Return articles **authored** by:
e.g., *"PJ Hayes"* or *McCarthy*

Return articles **published** in:
e.g., *J Biol Chem* or *Nature*

Return articles **dated** between: 2014 — 2016
e.g., *1996*

Several resulting articles mention the Geriatric Depression Scale and give a reference to a 1982 article published in *Clinical Gerontologist* or a 1983 article published in *Journal of Psychiatric Research*.

For more help with databases, search terms, and other library challenges, contact the embedded librarian for this course, Catherine Hendrickson, 909-389-3551 or email her at chendric@craftonhills.edu. The time to do this is the first week of the semester, not the week your project is due.

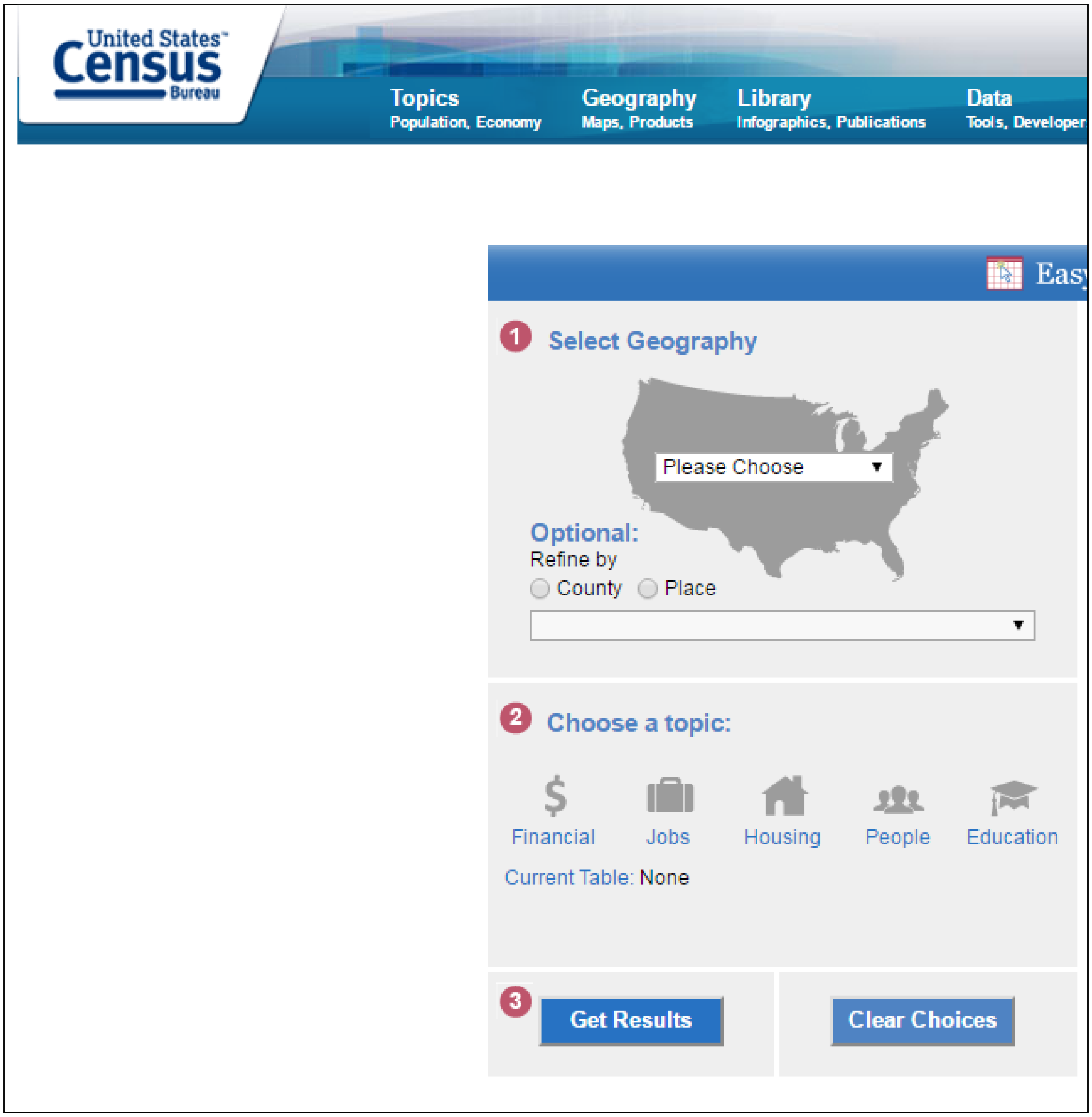
Complete Bibliographical Information

Make sure that you get complete bibliographical information for each article and record it in APA format. If you do not have all of this information, go back to a Scholar Google advanced search and type in what you do know about the article (e.g., authors' last names, year of publication) and it will find the rest of the information (e.g., title, journal name, volume, page numbers). Here is what we mean by complete bibliographical information.

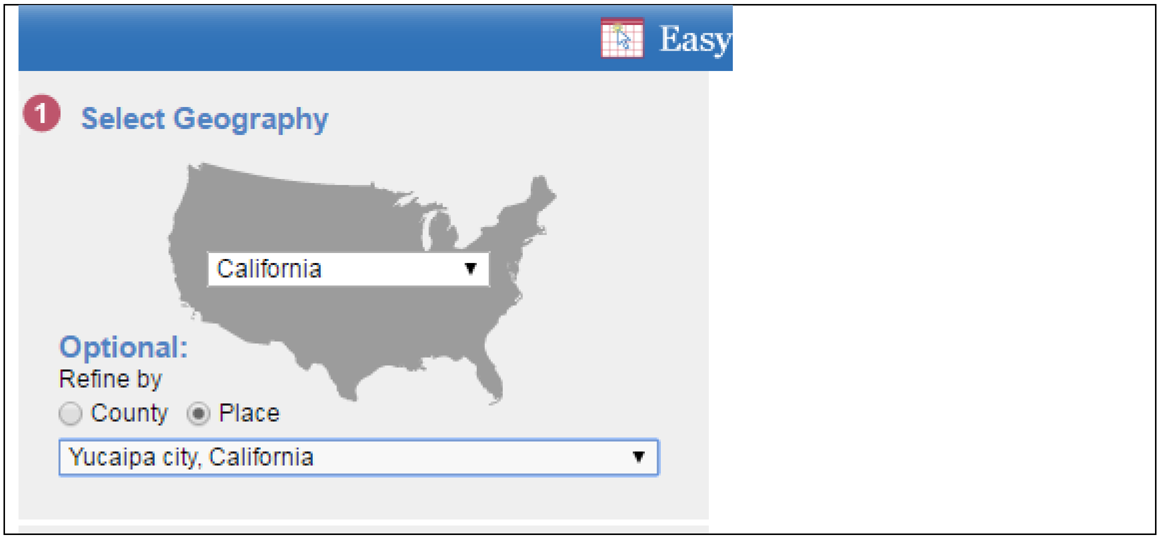
Book	Authors' names (last name first), Year, Title (italicized), City of publication, Publisher
Chapter in Book	Authors' names (last name first), Year, Title of article, in Title of Book (italicized), name(s) of editors, City of publication, Publisher, page numbers
Article in Journal	Authors' names (last name first), Year, Title of article, Name of Journal (italicized), volume number (italicized), page numbers
Article in Newspaper	Authors' names (last name first), Year, Title of article, Name of Newspaper, date, page numbers
Conference Presentation	Authors' names (last name first), Year, Title of presentation, Name of Organization, City of presentation, date
Website	Authors' names (last name first), Year (last revised or accessed), Title of article, URL

A *census* is information about an entire population. Background data on the U.S. Population can be found in the U.S. Census. Here are two easy ways to search it.

http://www.census.gov/easystats/#



So, I just use the drop down menu to select California and then click "place" and then I see another drop down menu and toward the bottom I select Yucaipa.



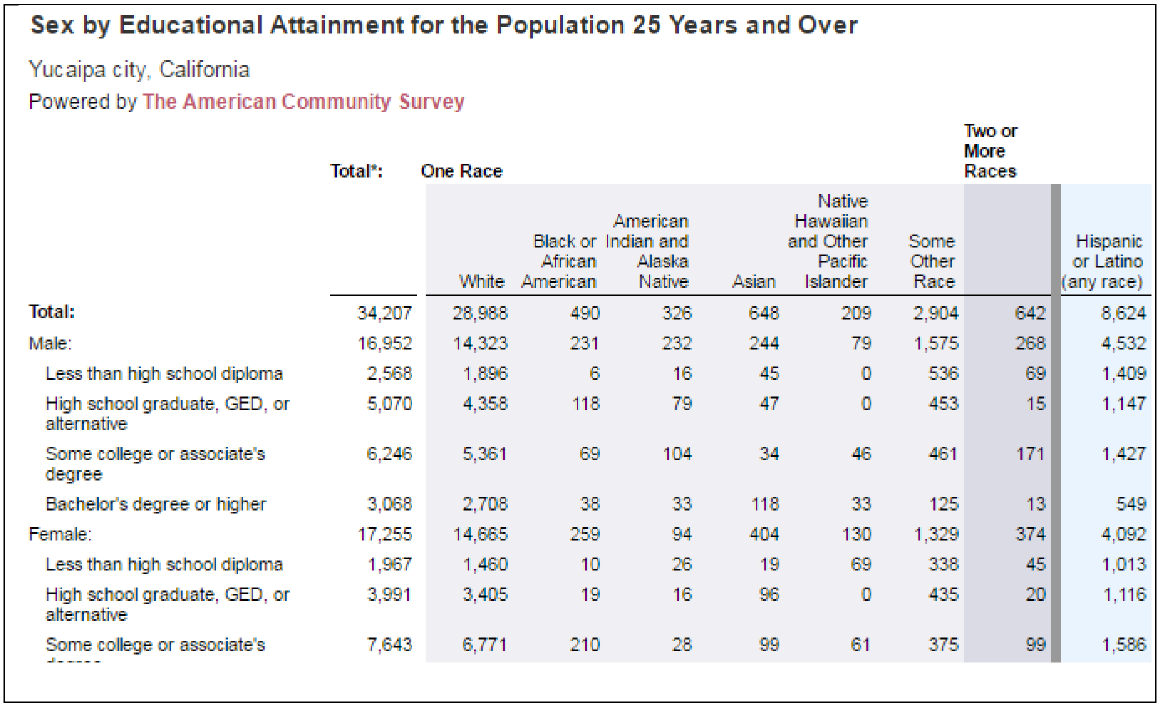
1 Select Geography

California ▼

Optional:
Refine by
○ County ● Place

Yucaipa city, California ▼

I then click on “education” and select college education for adults over age 25 and get this table.



Sex by Educational Attainment for the Population 25 Years and Over

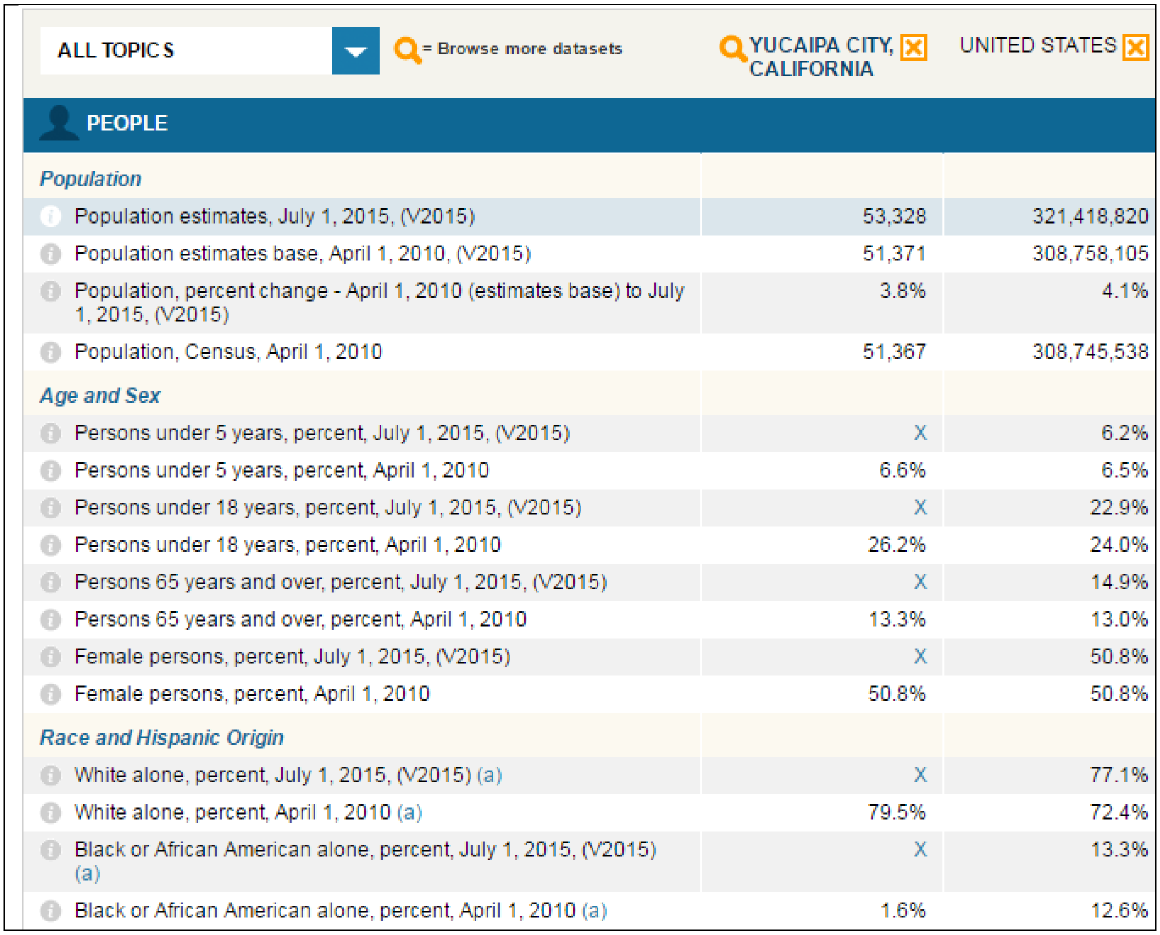
Yucaipa city, California
Powered by The American Community Survey

	Total*:	One Race						Two or More Races	
		White	Black or African American	American Indian and Alaska Native	Asian	Native Hawaiian and Other Pacific Islander	Some Other Race		Hispanic or Latino (any race)
Total:	34,207	28,988	490	326	648	209	2,904	642	8,624
Male:	16,952	14,323	231	232	244	79	1,575	268	4,532
Less than high school diploma	2,568	1,896	6	16	45	0	536	69	1,409
High school graduate, GED, or alternative	5,070	4,358	118	79	47	0	453	15	1,147
Some college or associate's degree	6,246	5,361	69	104	34	46	461	171	1,427
Bachelor's degree or higher	3,068	2,708	38	33	118	33	125	13	549
Female:	17,255	14,665	259	94	404	130	1,329	374	4,092
Less than high school diploma	1,967	1,460	10	26	19	69	338	45	1,013
High school graduate, GED, or alternative	3,991	3,405	19	16	96	0	435	20	1,116
Some college or associate's	7,643	6,771	210	28	99	61	375	99	1,586

These data about the population norms for Yucaipa can be used in several ways for your project. First, it may help set up a hypothesis. Second, it can be used in describing your sample or site at which the research was conducted. Third, if you have a hypothesis about your sample differing significantly from these norms, you can test with inferential statistics. Fourth, these data may be useful in your discussion section, attempting to explain your results.

Remember, you might have to do some calculations on these raw data to transform them into useful descriptives. Knowing that there are 648 Asians in Yucaipa is not as clear as the descriptive statistic of the part/whole percentage. To get to that, divide 648 (the part that is Asian) by 34,207 (the whole number of Yucaipa residents), then multiple the quotient by 100, yielding about 2%.

Here are demographic data for the U.S. (e.g., age, ethnicity) and for California. You can type in a specific local area and see the comparison.



PEOPLE	YUCAIPA CITY, CALIFORNIA	UNITED STATES
Population		
Population estimates, July 1, 2015, (V2015)	53,328	321,418,820
Population estimates base, April 1, 2010, (V2015)	51,371	308,758,105
Population, percent change - April 1, 2010 (estimates base) to July 1, 2015, (V2015)	3.8%	4.1%
Population, Census, April 1, 2010	51,367	308,745,538
Age and Sex		
Persons under 5 years, percent, July 1, 2015, (V2015)	X	6.2%
Persons under 5 years, percent, April 1, 2010	6.6%	6.5%
Persons under 18 years, percent, July 1, 2015, (V2015)	X	22.9%
Persons under 18 years, percent, April 1, 2010	26.2%	24.0%
Persons 65 years and over, percent, July 1, 2015, (V2015)	X	14.9%
Persons 65 years and over, percent, April 1, 2010	13.3%	13.0%
Female persons, percent, July 1, 2015, (V2015)	X	50.8%
Female persons, percent, April 1, 2010	50.8%	50.8%
Race and Hispanic Origin		
White alone, percent, July 1, 2015, (V2015) (a)	X	77.1%
White alone, percent, April 1, 2010 (a)	79.5%	72.4%
Black or African American alone, percent, July 1, 2015, (V2015) (a)	X	13.3%
Black or African American alone, percent, April 1, 2010 (a)	1.6%	12.6%

Yucaipa is fairly representative of national norms in terms of age or gender distribution, but less so in its ethnic distribution.

To get demographic data on ethnicity, gender, age, education and income levels on a state, city, and neighborhood level, use City-Data, as demonstrated by this video showing how to navigate around that site.

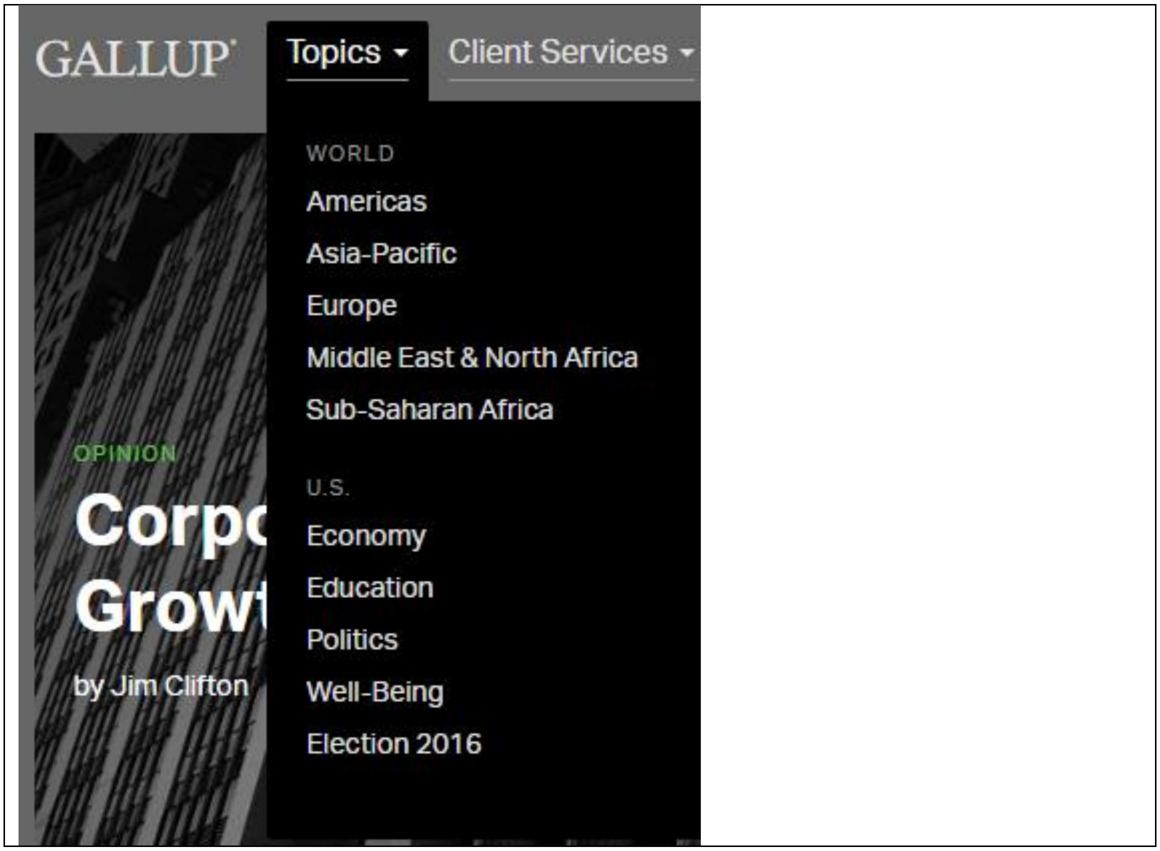
Another website with neighborhood level demographic data (as well as crime, housing prices, and schools) is Trulia, as demonstrated by this video showing how to navigate around that.

Polling Data

Polling data (and some background data) can be found at the General Social Survey maintained by the National Opinion Research Center of the University of Chicago. Getting these data requires a bit of work.

The annual study of entering freshman is conducted by a center at UCLA. Here are some brief infographics for recent data. Here is a recent complete report. This .pdf file can be search with a control F, and then type in a search word such as “religio” and it will go to 67 places in the document where there are words like religion, religious, religiosity, etc. Some of these appear in the discussion of the findings, but you can also find tables of the religious preference of college freshmen.

The best private polling organization in the U.S. is Gallup Click on “topics” and you will see quite an array dealing with political, economic, and personal questions.



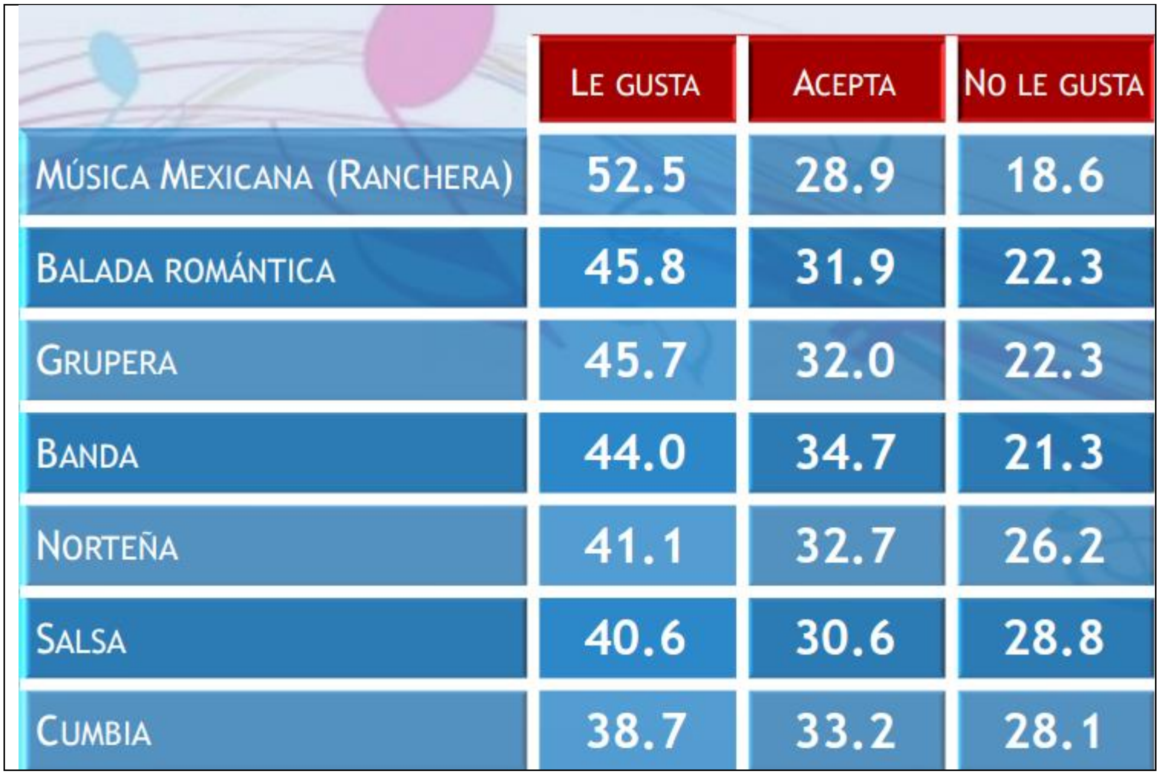
Perhaps the one most relevant to psychology would be “well being” so let’s click on that. Now I see an array of current topics, perhaps on health care or food insecurity. One topic on millennial workers catches my interest, so I click on that.

The data can be used in my literature review to set up my hypotheses (or in a sample vs. norms design). The analysis by Gallup’s authors can be used in my introduction or discussion. The greatest use of these polls would be methodology. Gallup knows how to phrase questions and response formats. Don’t try to come up with your own questions when Gallup has already developed some and field tested them on a sample of thousands.

There are other polling organizations in the U.S., such as Roper and People-Press. Zogby is one of the best about the

Middle East. Barna specializes with Evangelicals, but they also have studies about Millennials in the workplace. Polling Report summarizes many other polling organizations.

In Mexico, the best polling is done by Mitofsky. You could compare the attitudes of persons of Mexican descent living in the U.S. with the national attitudes in Mexico on topics such as homosexuality, levels of stress, cancer attitudes, bullying or musical preferences.

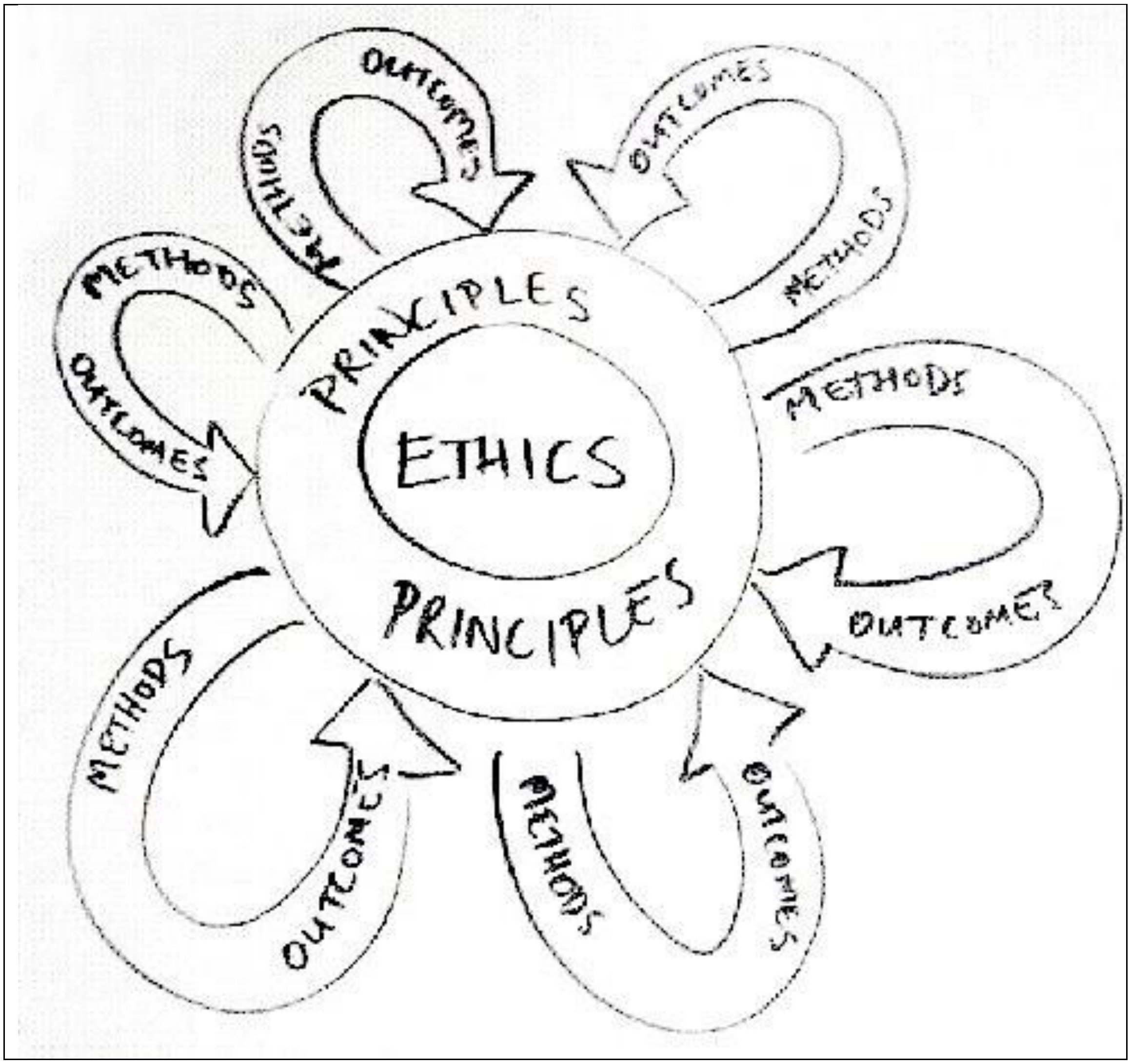


	Le gusta	Acepta	No le gusta
Música Mexicana (Ranchera)	52.5	28.9	18.6
Balada romántica	45.8	31.9	22.3
Grupera	45.7	32.0	22.3
Banda	44.0	34.7	21.3
Norteña	41.1	32.7	26.2
Salsa	40.6	30.6	28.8
Cumbia	38.7	33.2	28.1

P.S. When you come to my house, we dance cumbia (and reggaeton). OK, that’s your fun, musical break. Now, get to the library and start your secondary research.

Chapter #3: Research Ethics

The study of *ethics* concerns the reasons why a given action is to be regarded as morally right or wrong. Ethics is a branch of philosophy, and there are applications for the practice of business, medicine, government, and scientific research. Ethical considerations apply both to the actions we take as well as the impact of our actions on others. For researchers, this means that we have to be aware of the ethical implication of our methods and our outcomes.



In psychology, the science of behavior, the organization articulating research ethics is the American Psychological Association (*APA*). The site summarizes APA ethical principles for researchers.

Government agencies, such as the National Institutes of Health (*NIH*) and National Science Foundation (*NSF*) often enforce ethical principles in terms of guidelines for funding. Several of these organizations have ethical training and certification procedures. One of the most widely used is the Collaborative Institutional Training Initiative (*CITI*). Many universities (e.g., Loma Linda) require that their research faculty be certified by CITI or NIH.

Duty to Science

One of the ethical principles of all scientific research concerns the researcher's duty to science (and all those persons who will someday act based upon assumptions of the validity of reported research). The term for fabricating raw data is *dry labbing* (which comes from trying to do a chemical experiment without any test tubes). It is also obviously unethical to intentionally distort the collection of data either by intentionally changing the numbers to fit a favored hypothesis or by eliminating some cases just because they don't fit that hypothesis.

Less obvious, but perhaps more common than the aforementioned types of fraud, is the process of adding new cases, while keeping a running analysis of the data, and then stopping when the data become statistically significant. This is sort of like flipping a coin over and over again until you get a run of three heads, and then stopping and saying that you have proved that the coin has a tendency to come up heads. It is abusing naturally occurring chance variations to try to claim that the results were not produced by chance. It is similar to the *confirmation bias* we see underlying most *common sense* and *pseudoscience* explanations of behavior and mental processes: just noticing those cases that happen to fit in with the theory (and are therefore seen as cases that confirm the theory).

Ideally, the way to deal with this problem would be to state in advance how many subjects will be run and how many repeated measures will be made. There are even websites where researchers can pre-register their forthcoming studies so as to make a public commitment to adhere to a specific number of subjects or trials. One such site is the Open Science Framework (OSF). More suggestions on transparency can be found at the Berkeley Initiative for Transparency in the Social Sciences, (*BITSS).*

The move to transparency and openness promotion (*TOP*) has sparked some important statements that hundreds of journals and organizations have endorsed. The APA has endorsed many of these data sharing principles. The *APS* (which used to stand for American Psychological Society but now stands for Association for Psychological Science) even awards badges to researchers for their adherence to the principles of open data, open materials, and pre-registration.

That is why in this class you will be required to have your data on a Google sheet and your questionnaire as an appendix of your write-up on a Google doc. (While we do not enforce pre-registration or commit to a maximum sample size, we will not begin statistical calculations until the data gathering has ended.)

This is because, in practice, we usually don’t know how many participants we will be able to secure. In such cases of opportunity and snowball sampling, we should not be running an ongoing analysis of the data, consciously seeking just a few extra subjects to try to put us into the significant range of p values.

A related challenge to science is *data dredging* or p hacking, data snooping, backtest overfitting, torturing the data until it admits to some significant relationship about something. This is enabled by an extremely large number of variables (which becomes very likely in the case of the big data generated by personal digital devices or centralized medical records). With a thousand correlations, we should end up with fifty of them significant at the $p < .05$ level by pure chance alone. So, again, the problem is that we are abusing naturally occurring chance variations to try to claim that the results are significant (i.e., not produced by pure chance). There are many great examples of such meaningless (though strong) correlations at Tyler Vigen’s website.

Less obvious than conscious attempts at data dredging would be the *file drawer problem*. This is another problem due to large numbers of studies producing a few that look significant (statistically) when there is no real causal relationship between the variables. But in this situation, the problem is not due to intentional actions of individuals, but to the policies of institutions (especially scholarly journals).

Here is an example. Suppose that across the country there are a thousand departments of psychology doing a study on mental telepathy (which most scientists regarded as a pseudoscience). Let’s say that each of these thousand departments does a well-designed study of the topic. Let’s say that mental telepathy does not exist, so subject performance is only due to luck, and therefore only fifty of these thousand studies will show significant results at the .05 level. Unfortunately, it will be these fifty studies that are sent off to journals (and will be most likely to be published because the topic is important, and we said that the studies were well designed, and now we see significant results). So, for the year there could be fifty well designed, statistically significant, published studies confirming mental telepathy. The problem is that the preponderance of evidence (the other 950 studies) showed that chance was a better explanation for the data, but the data from those studies remain at the bottom of some "file drawer" and are not published.

One of the solutions already mentioned (pre-registering upcoming studies) is one possible approach insofar as it gets those other studies (lacking significance) out into public view. Another solution is to encourage more replications of such incredible findings (and publish those replications even when they don't achieve significance).

Perhaps the best way to preserve fairness with science is to continue to insist on rigorous *peer review***.** Just as experts can detect sloppy methodology and analysis, it is now possible to use algorithms to identify some of the aforementioned forms of fraud.

The Office of Research Integrity operates under the U.S. Department of Health & Human Services. The *ORI* investigates violations of integrity, providing assistance to institutions responding to charges of investigator misconduct involving Public Health Service funds. When violations are confirmed, the guilty parties are publicly identified and retractions of published research are demanded. Over a hundred cases of research misconduct are reported each year to this agency alone. Perhaps its most useful function is that this agency provides training in how to comply with ethical guidelines and how to respond to misconduct. Here is a great (but long) interactive video showing a dramatized example.

Duty to Other Scientists

Another ethical principle is that of academic integrity. Each scholar has a duty to fellow scholars. One obvious rule is against *plagiarism*: when you use someone else's words as if they were your own. If you are intentionally and directly quoting a sequence of more than a few words, you need to put those words in quotation marks and indicate who said or wrote them. Even if you change a few words in a paragraph, but leave most of them (as in Mrs. Trump's reworking of Mrs. Obama's convention speech) you should indicate from whom you are quoting.

But the duty to one's fellow scientists goes beyond avoidance of the outright plagiarism described above. If data or a theory have their origins in some other researcher(s), we should acknowledge that role by making a formal *citation*. This does not diminish the scholarship of your own work, but increases its status by showing that your literature review is more comprehensive. (Many articles are rejected for publication because peer reviewers note a lack of citations to previously published research and theory.)

Another duty to our fellow scientists specifically applies to colleagues who helped us in developing our research. We should acknowledge those colleagues who helped us with the data gathering, analysis, or writing of our finished research report. The criteria about who should be included as co-authors should be set forth before an article is written or before an abstract is submitted for presentation. In general, someone who has performed merely clerical assistance (e.g., handing out questionnaires, coding data into a spreadsheet) would not qualify as a co-author. Actually participating in the analysis (e.g., formulation of hypotheses, interpretation of the results) is usually required for co-authorship. Some journals now require a footnote where the specific role of each co-author is clarified.

This raises an ethical concern about the reverse situation. Suppose an untenured researcher, Dr. Y, writes an article without the help of her supervisor, Dean X, a vain administrator who is envious of the ability of the scientists under him to publish their research. Dr. Y hopes to put herself in good standing with Dean X by naming him as a co-author. This is similar to giving an unworthy student a higher grade than that which was actually earned. There is no direct or immediate harm to the individual perpetrators of this fraud, but the real victims are the overall scientific community who are deceived into believing that Dean X is a competent researcher, and the larger result is that the merit accrued to the real authors of all published research articles is somewhat diminished.

Yet another ethical concern is a possible *conflict of interest* (political or economic) that a researcher might have. This holds whether the interest is just with the individual researcher or whether it involves the institution supporting the research. The most obvious situation would be where a pharmaceutical company is funding psychiatric research on the effectiveness of its product. The researchers should report this source of funding in their presentation of the findings (especially a published article).

Duty to Animal Subjects

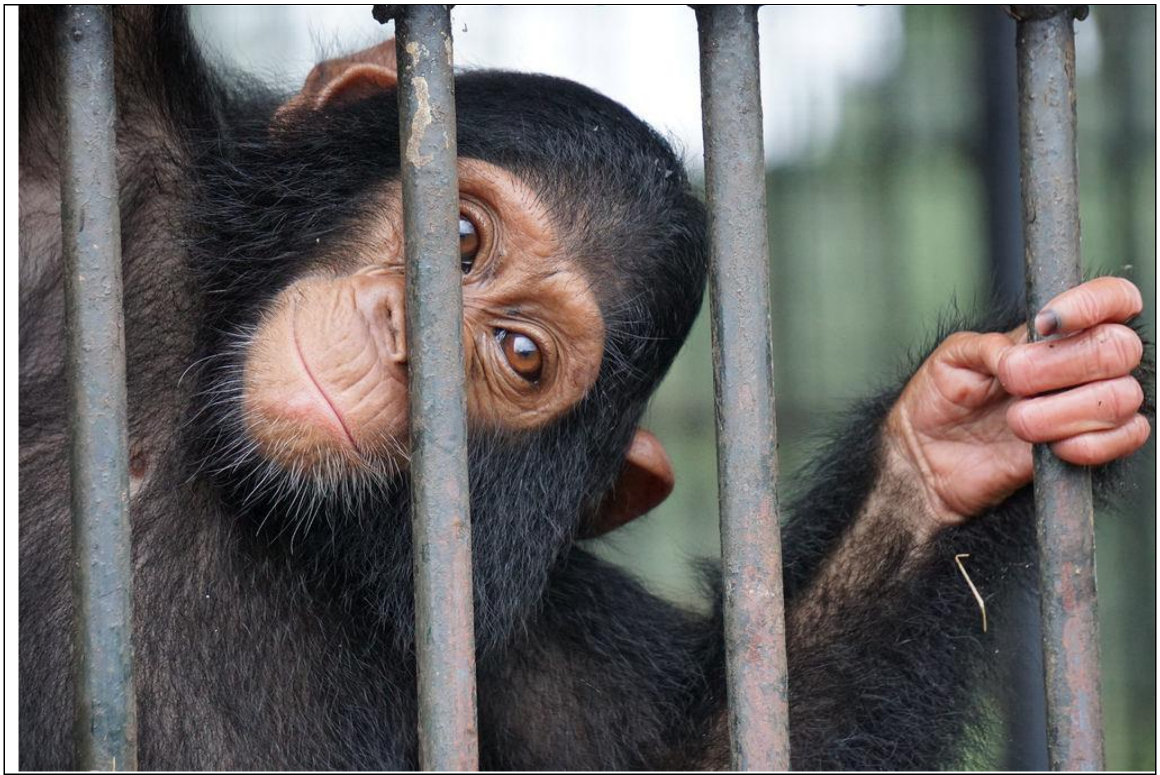
Doing psychological research with humans is not always possible. It is difficult to keep them in laboratory conditions 24/7 throughout major portions of their lifespans. It is obviously unethical to deprive humans of certain experiences (e.g., parental love, education) during their formative years, or to subject them to life-threatening conditions. That is why so much research (medical and psychological) is performed on non-human species. These would include some of the greatest studies within psychology: Pavlov and his dogs, Watson and rats, Skinner and pigeons, Lorenz and the ducklings, Harlow and the monkeys. Each year nearly a million non-human animals (mostly rabbits, rats, Guinea pigs, and primates are used in biomedical research). That number has finally leveled off and may be declining.

If you were on an IRB, using today's ethical standards, would you support Harlow's research?

Research subjects, human and animal, need to be protected against some of the possible dangers of medical and psychological research. The APA's emphasis on ethics in psychological research goes back to the 1920s with the Committee on Animal Research Ethics (CARE). Now each institution conducting research on animals is supposed to have an Institutional Animal Care and Use Committee (*IACUC*). Such a board is composed of at least five members, one of whom should be a veterinarian.

The IACUC must be satisfied that the researchers are taking proper precautions with respect to the animals' housing, nutrition and health care. A frequent sticking point is the design of the cages or the level of cleanliness to be maintained. A main point of review must be whether the degree of pain and suffering experienced by the animals is justified by the purposes of the research. A study designed to cure cancer may justify more risk to the animal subjects than research testing cosmetics.

What varies greatly is how different species are treated. Invertebrates may receive very little attention, while pigeons and rodents (e.g., rats and Guinea pigs) will receive more. The highest level of concern is usually reserved for primates, based upon the assumption that these creatures are capable of some degree of reflection and emotional suffering. Accordingly, the NIH announced in December 2015 that it would no longer fund chimpanzee research and it would retire its own chimp research subjects.



NIH is ending funding of research on chimpanzees.

Duty to Human Subjects

The APA's emphasis on ethics in human subjects research goes to the 1920s with the Committee on Human Research (CHR). Interest in ethics was heightened after World War II, when the *Nuremburg* Trials examined some of the research done by Nazi scientists in the death camps of the Holocaust.

The avoidance of unnecessary *risk* to the research subjects is even more important when the organisms are human. Such risk could be in the form of death, pain, discomfort, injury, or (when it comes to human subjects) emotional distress from embarrassment and guilt.

Unlike animals, human subjects
can experience embarrassment.

Researchers have a duty to preserve and protect the well-being of subjects, and to justify any risk in terms of likely benefits of that research. With human subjects, the benefits must also be likely for the individual subject involved. Therefore, we could not justify Harlow's study on maternal deprivation (conducted on motherless monkeys) for human infants. Clearly, the *Tuskegee* syphilis study was unethical since the men serving as subjects only bore the risks of increased disease, disability and death, and received none of the benefits of whatever future syphilis treatments might be developed from the research.

Oliver Wenger, architect of the Tuskegee Study

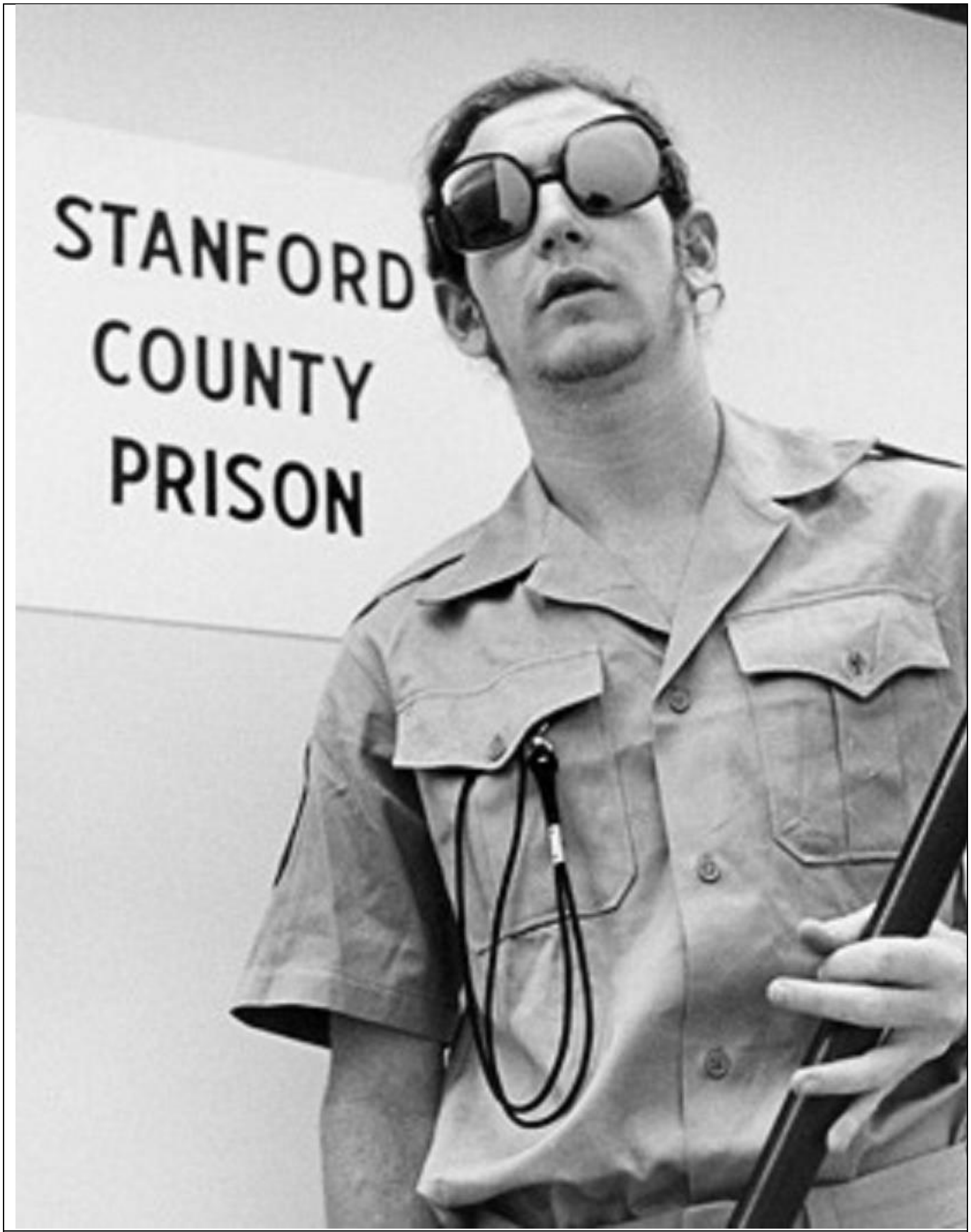
Another ethical principle for human participants would be respect for their privacy. Identifying personal information should not be disclosed. These concerns are written into federal law: Health Insurance Portability and Accountability Act (*HIPAA*) for health care patient records and Family Educational Rights and Privacy Act (*FERPA*) for education records of students. *Anonymity* refers to the identity of the subjects being unknown to the researcher. *Confidentiality* refers to the situation where the identity of the subjects, though known to the researcher, will not be reported to someone listening to the conference presentation or reading the article reporting the research.

Another set of ethical principles revolves around the concept of *informed consent.* This means that the subjects must be true volunteers, and not coerced into participating. Student subjects should not be bribed with the prospect of a higher grade for participating (unless those who decline are given some alternate route of earning the credit). Incarcerated subjects should not be offered promises of a reduced sentence for participating. Access to public funds (e.g., welfare, pensions, section eight housing, food stamps, SNAP) should not be conditioned on participation in research. Desperately impoverished individuals should not be given irresistible financial incentives. Subjects should be allowed to decline participation at the outset, or withdraw from the research after starting, with no fear of retaliation from the researchers.

One complication affecting informed consent is that sometimes a degree of *deception* is required in order to measure the subject’s response to certain situations, especially in the branch of social psychology. We can’t just ask if a person is prejudiced against other ethnicities. We may find it convenient for our subjects to believe that we are really doing research on something else, and see if the ethnicity of a stimulus person makes any difference in the subject’s behavior.

In many studies, deception takes the form of the use of a *confederate,* a person whom the subject assumes to be just another participant in the research, or even a bystander, but certainly not as someone playing a role designed by the researcher. For example, in Asch’s study of conformity, the five other young men who intentionally gave the wrong answers about the length of the lines would be confederates. The subject had to think that the confederates were giving their own estimates in order to see how the subjects would be influenced by other people. In Milgram’s study of the willingness to give electric shocks, the actor who pretended to receive the shocks would be a confederate (and another confederate would be the man in the white coat with the clipboard giving the instructions to shock the person in the next room). If these actors had not played their parts and the subjects knew that the person in the next room was not really getting the shocks, we could not have investigated the willingness to give those shocks. When such a serious degree of deception takes place, subjects should definitely be *debriefed* after their responses have been measured so that any residual guilt, embarrassment or other emotional trauma can be alleviated.

Another problem with informed consent is how certain individuals might not be capable of giving it. The most interesting topics in psychology (e.g., autism, dementia, schizophrenia, Down syndrome, infant development) involve research done on persons who (by virtue of the fact that they have the aforementioned conditions) cannot give informed consent. Any time we are doing research with minors, the developmentally disabled, or those with delusional mental illness, or dementia, we have to confront the lack of informed consent. In some cases it would be necessary for the responsible adult charged with making that person’s medical decisions (a legal guardian, conservator, or site manager) to give formal consent for participation.



Phil Zimbardo's prestige within psychology is not due to the brilliance of his design of the Stanford Prison Study, but to his ethical response in halting it when it became obvious that it was becoming harmful to the subjects.

It is usually not sufficient for an individual researcher or team of researchers interested in a given project to be the sole source of ethical review and approval. The applicability of the guidelines to a specific case may be unclear. In most research institutions there is an internal institutional review board (*IRB*) or independent ethics committee (IEC) that must approve scientific research projects involving human subjects.

Certain studies are more easily approved on ethical dimensions. For example, institutions have a right to gather data: employers gather data on their workers and customers, hospitals gather data on their patients, schools gather data on their students. These data are exempt from ethical review if the data are normally gathered and reflect standard sorts of activities for that organization. There would be no need to get a signed consent form from each participant in such an archival study. However, if you are only a student, intern, volunteer, or employee of such an organization, do not assume that you have the right to access or permission to use the data in the archives. You need to find out organizational policy (especially if there are FERPA or HIPAA applications) and get the approval of someone in authority. Hint: have a meeting with the director and present your proposal. One element of your data gathering might be unacceptable, and you could be able to negotiate an alternative study by changing a few variables or procedures. Obviously, you are responsible for maintaining the subjects' confidentiality in the reporting of the data, but the organization may also want to use additional safeguards for preserving the anonymity of the subjects.

Similarly, a field count based upon public behavior (e.g., shoppers walking into a store) where there is no expectation of privacy could be conducted without formally obtaining consent forms. (But, publically releasing photos or recordings of specific subjects could be a violation of confidentiality.)

If the data are in narrative form (e.g., the subjects' own words from an interview, focus group, a post on a threaded discussion board) we would have to take a look at the right to assume a private conversation. I would impose a higher standard of protecting confidentiality for the case study of a patient in psychotherapy, or an interview with a survivor of sexual abuse, or a minor, and a lower standard of confidentiality for someone like a politician making a public speech. The duty to protect the subject's identity would be greater in the case of subjects who would be vulnerable to the release of data about sensitive topics (e.g., whistleblowing, drug use, financial information, immoral / illegal behavior).

Subject or type of Research	Responsibility to protect patient identity
Minor in age	Great
Psychiatric Patient	Great
Student's grades	Great
Sensitive topic	Great
Embarrassing behavior	Great
Adult subjects' attitudes about products or politics	Moderate
Adult behavior in public	Moderate
Historical figure	None
Public figure's public statements	None

Some IRBs have a policy of completely exempting review of studies where the variables and methods are routinely measured if the topics are not sensitive, there are no known mental, physical, or economic risks, and the population is not vulnerable. IRBs may have expedited reviews of studies involving minimal risks such as moderate exercise or stress from testing and surveys, as long as the population not vulnerable. Full IRB review is called for when there are questions about sensitive topics (e.g., criminal activity) or heightened stress produced by the measures (e.g., strenuous exercise, frightening situations) or vulnerable populations (e.g., minors, elders, prisoners, in patients, pregnant women).

Chapter #4: the Proposal

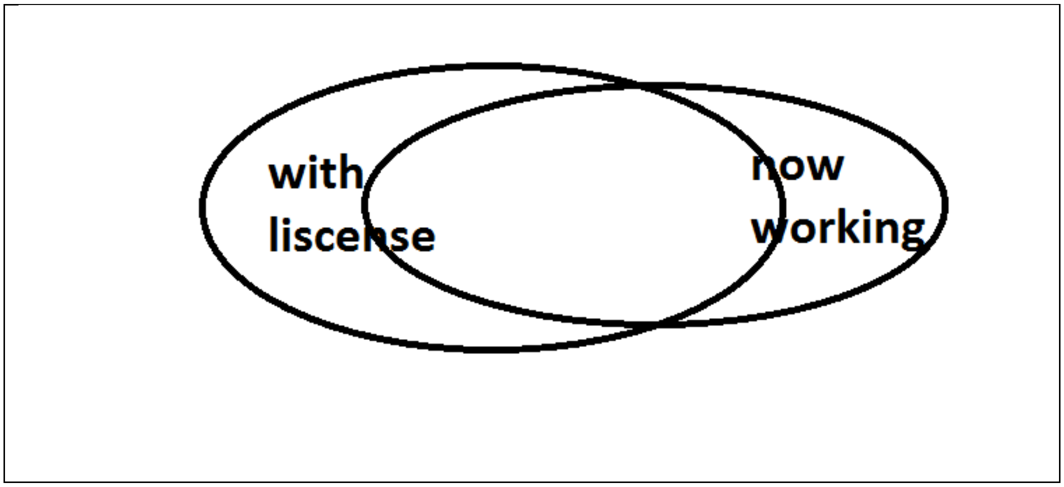
Psychology is defined as the scientific study of behavior (and mental processes) in humans and animals. Humans and/or animals are therefore the *organisms* from which psychologists must get their data. Other terms for describing a single organism would be a case, *subject* or *participant*. Most psychology journals now prefer to use the term “participant” but in this class we may use organism, case, or subject interchangeably. Notice that the term “subject” always refers to a person or animal being studied, not the topic of the study. That *topic* is some aspect of the subject’s behavior, a dependent variable, usually the criterion variable.

Population, Sample or Group

A *population* is the type of organisms studied in the research. Here are some examples of populations.

- All Fisher 232 rats
- All residents of the United States
- All potential voters in the U.S. election
- All students at Crafton Hills College
- All living veterans who served in the U.S. military
- All women who have suffered from domestic violence
- All customers who have purchased automobiles in 2017

- All currently licensed truck drivers
- All individuals currently employed as truck drivers



Notice that slight variations of phrasing redefines a population. All potential voters in the U.S. election would be a smaller circle, inside the larger circle of all U.S. residents (since not all residents vote). The last two populations would be overlapping circles because not everyone with a license to drive a truck still does so, and not everyone driving a truck for money is legally licensed to do so. In general, the more clearly defined the population is, the better job we do in controlling confounding variables.

Unless we are doing a complete *census*, we will not hope to obtain data from every member of the population. We will only obtain data from some members of the population, and these subjects actually observed (from whom we obtain data) are known as the *sample*. The two criteria for a good sample include being large and being representative of the population.

Sampling is the process of selecting specific subjects from the population to form our sample. It would be a poorly done sampling (and unethical science) to intentionally select only those subjects who would be most likely to support a hypothesis. Unnecessarily small samples are also less likely to be representative of the population (and more likely to demonstrate a trend that is not really statistically significant). A truly *random* sample is one where each member of the population had an equal chance (compared to every other member of the population) to be selected for the sample. For example, if the population was a thousand, and the sample size was a hundred, if each and every member of the sample had a 10% chance of being selected for the sample, we could say that the sample was truly random. "Random" should not imply that the sampling was done in a haphazard fashion. One way to approach randomness would be to use fair methods of chance for selection, such as flipping a coin, lotteries, etc.

In practice, truly random processes may be difficult to implement. A frequently used alternative would be *stratified* samples, which intentionally strive to select a proportionate amount of each sub-population. For example, if the registered voters of a given congressional district are 52% female, and we want a sample of 1,000 in order to conduct a poll, we would make sure and select 520 women and 480 men. If that district's registration was 42% Democrats, 38% Republicans, and 20% other; we should have those proportions within our sample as well. *Quota* sampling is similar to this, but only a minimum number for each sub-population is established, not an exact proportion, as in a stratified sample.

A *cluster* sample is where we select (randomly or because of proportionate representation) a specific subset of the population from which to draw our sample. For example, a study of the population of university students might pick just a few representative (e.g., demographically diverse) campuses on which to distribute the questionnaires. A study of the population of voters in a congressional district might pick a handful of swing precincts, and try to get as many voters as possible within those clusters. A study of Starbucks customers might pick just a few randomly selected locations, and try to get as many customers as possible in each of those to fill out questionnaires.

Most student projects don't come close to being random, or stratified, quota, or even cluster. Here's what happens most of the time. You go to a location around campus with heavy traffic (e.g., outside of the library, on a quad area where there is a lot of foot traffic, a student union or major cafeteria. (At Crafton Hills College, the prime location would be under the breezeway of the building with the 39 steps, and the best time would be Monday – Thursday between 8:50 AM and 2:00 PM). Individually approach the students saying "Would you be willing to help me with a project I'm doing for a psychology class?" About half of the students will decline, saying "Sorry, I've got to go to class." (It helps if the instructor has a reputation for interesting (and short!) questionnaires. (So, definitely mention Brink's name.)

Such samples are called samples of *convenience.* If they are large enough and if you chose a location and time that attracts a diversity of students, your sample could still be pretty representative of the overall college population. However, choosing the wrong time and place could make the sample very *biased* (i.e., lop-sided in that it will overly represent one segment of the population and under-represent others). For example, if you distributed the survey on Friday morning only on the east end of campus, you would get mostly fire students (and mostly males) but if you distributed that survey on the west end of campus you would get mostly child development students (and mostly females).

Another problem in biased sampling comes when the sample is *self-selected*. If subjects had to make an effort to be included in the sample (e.g., click on a link, return a questionnaire in the mail) the subjects who were more passionate about the topic would be more likely to participate. That might explain why such internet polls (or those where subjects have to send a text, or call an 800 number) would be more likely to attract those who really care about the topic. Many such internet polls in the 2016 election showed Sanders and Johnson far ahead of Clinton (who tended to attract lukewarm support, at best). These samples tend not to be representative of the larger population and therefore, not very good predictors of the results on election day (when a lot of lukewarm voters finally make up their minds and cast a ballot).

Once we have obtained our sample, we may decide to divide it up into separate *groups* and then compare those groups on some dependent variable. In an experiment, these groups are *randomly* assigned, and treated differently with respect to an independent variable. In other forms of psychological research (e.g., quasi-experiment, survey, correlational) the grouping is determined by background variables or even by the subjects' own preferences.

This video clarifies the terms: subject, population, sample, group. Do not call everything a group, and do not throw the terms together (e.g., "sample population" or "population group" or sample group").

A *constant* is a measure that does not change within a sample (or we could describe a constant for a population or group). A variable is a measure changing from one subject to another. For example, if all members of our sample were male, gender would be a constant. If both males and females are in the sample, then that is a variable that we should measure, as further explained in this video about variables.

Within the group of males, gender is a constant (as it is within the group of females) but within our entire sample of males and females, gender is a variable to be measured.

Four Ways of Dealing with a Variable

In your research proposal, there must be at least one variable (usually, a dependent variable) that each subject within the sample is measured on. (If there are only constants, we cannot test any hypotheses). How that variable is actually measured (e.g., categories, levels, ranks, numbers) constitutes the *operational definition* of that variable.

Dependent variables must be measured, but independent variables can be dealt with in any of four ways. First, we can measure them on some quantitative scale (then correlate them to the dependent variable). Second, with independent variables that are stimuli, the researcher may be able to figure out how to *manipulate* the variable. That means that the researcher forces some subjects to have a high level of the independent variable and other subjects to have a lower level of the independent variable. In psychiatry, when some patients (i.e., the participants) get the real medication (while the other group only gets a *placebo*) that is manipulation of the independent variable of treatment. In industrial psychology, when some of the workers (i.e., the participants) are assigned to get a new form of training, that is manipulation of the independent variable of training.

Manipulation of an independent variable is what makes research an *experiment* instead of a survey. Notice that it is not manipulation (and therefore not an experiment) if the researcher merely measures the independent variable. If I measure a person’s age, that is not manipulation of the variable of age. If I ask if a worker has received some training, that is not manipulation of the variable because whether or not the worker was trained in the past would be based upon someone's decision (perhaps the worker himself or his boss) and not an assignment by the researcher. This video gives other examples of experiments versus other forms of observational research.

Suppose we are not really interested in the impact of a particular independent variable (e.g., age, gender, ethnicity, religious upbringing) because our topic is the role of training (the independent variable we want to study) on worker performance (the dependent variable we want to study). We should still concern ourselves with the possible impact of those independent variables on the dependent variables, because they could distort our interpretation of the causal relationship. Such variables are known as *lurking* or *confounding* variables, as explained in this video.

Fortunately, there are two more ways to deal with independent variables to prevent them from becoming confounding. The third way to deal with an independent variable is to control it, and take it out of the picture. The easiest way to *control* a variable is to change it into a constant. This can be accomplished by exclusive sampling. Suppose you think that gender might impact worker performance on an assembly line task because women have better fine motor skills compared to men. If you simply measured whether or not a worker had received special training (and not the gender of those workers), this could be a confounding variable because maybe women were more likely than men to sign up for the training. So, you measured (and did not manipulate the variable of training), but did not measure or manipulate gender. Gender could therefore account for the difference between the trained and untrained workers, unless you controlled for gender by only including women in your sample. Now you are comparing "apples to apples" or more specifically, untrained women to trained women (and not mostly female trained workers to mostly male untrained workers).

Another way to control for gender would be to proportionately represent each gender in each of the training groupings. For example, if the population of assembly line workers at this company was 67% female, then we should strive to have something close to two-thirds of our trained workers and two-thirds of our untrained workers be females. The more we allow the two groups (trained vs. untrained) to differ in terms of their percentage of females, the more we are introducing a potentially confounding variable.

The fourth way that an independent variable can be dealt with is via *randomization*. The term random means equal probability. We can only justify our claim that a sample has been randomly selected from the population if each member of that population has been given an equal chance to be chosen to participate. Unless we have taken extraordinary means to assure this, we should not call a sample "random" but admit that it was selected out of the researcher's convenience (i.e., who was willing and available to participate). The more that an individual in the population has the ability to self-select into participation (or opt out) the less truly random the sample is. Surveys that solicit participation via mail or invitations to click on a link are self-selected and not randomly selected. This lack of random selection introduces confounding variables into the study, especially in a sample vs. norms design.

Another way that an independent variable can be randomized is through *random assignment*. This pertains to separate group designs, such as the previous example of trained and untrained workers. If we go on what workers themselves have chosen to do via training, we open up all kinds of other variables to influence their performance (e.g., gender, age, motivation). Random assignment means that each subject in the sample has an equal chance (compared to every other member of the sample) to be assigned to the treatment group. If the treatment group and sample are large enough, then we can safely assume that all these background and motivational differences have been equalized by random assignment. Each assigned group should be equivalent in terms of gender, age, ethnicity, motivation, childhood experiences, etc.

How to handle an independent variable		
	What this means	Difficulty
Measure	Use a scale: nominal, ordinal, interval, ratio	Some potentially confounding variables are hard to measure
Manipulate	Experimenter chooses different treatments for each group	Cannot manipulate background factors or subject choices
Control	Selective sampling to make this a constant; intentionally proportionate representation in each group	May not be able to assign subjects to groups
Randomize	Equal chance of being selected into the sample; equal chance of being assigned to the treatment group	May not have easy access to many subjects from which to randomly select; may not be able to assign subjects to groups in a truly random fashion

Four Designs for Testing Hypotheses

Whenever we test a hypothesis, we must do so in at least one of four ways: sample vs. norms, correlational, separate groups, or repeated measures.

Design	What this involves	Advantages	Potential problems
Sample vs. norms (sample vs. population; one sample)	Compare the entire sample to some external norms (usually coming from the population, or from a truly random distribution of the target variable)	Only have to measure one variable, one time; easy for field counts and archives when population norms exist	All the factors that made these subjects easier to sample (e.g., students, living in California, age) are confounding the results. When the sample was taken is another confounding variable.
Separate groups (independent samples; cross sectional, between subjects experiments; quasi-experiments)	Dividing the sample into separate groups and comparing the groups on the criterion variable.	Only have to measure (or manipulate) the grouping and measure the dependent variable; external norms unnecessary because we have a control group for comparison	Unless grouping was randomly assigned, any factors which influenced the grouping are confounding the results (e.g., cohort differences in cross sectional studies of aging)
Repeated measures (e.g., before & after, matched pairs, longitudinal, dependent samples, multiple evaluations, different aspects, within subjects)	Measuring the criterion variable more than once; requires coding of each subject's data	External norms and comparison groups are unnecessary because each subject serves as his own control; permits a smaller sample size	Anything that varies with time confounds: familiarity with previous exposure, practice, fatigue, boredom, the natural course of a disorder, Hawthorne effect, attrition

Design	What this involves	Advantages	Potential problems
Correlational, post hoc	Measure two variables quantitatively, then correlate them	No need for external norms, groupings, or repeated measures	Many correlations are spurious, leading to post hoc fallacies

Sample vs. Norms

The easiest approach is to compare our entire sample (on some variable) to some external norms. This is sometimes called a *one sample* design. Because the norms often come from population figures (e.g., a census) this is sometimes called sample vs. population. For example, we know that the population of the city of Redlands is 51% female. Our sample is shoppers (n = 50) going into the Redlands Sewing Center on a Friday morning. The variable on which we will compare sample and population with this field count is gender. We observe 46 women walk into the store. That observed frequency of 92% female in our sample is significantly higher than the norm of 51% for the city's population (p < .001). Therefore, we reject the null hypothesis and conclude that sewing customers (at least those observed in this sample) are disproportionately female.

Here is another example using the norms of assuming equal probability of random outcome. Subjects were inpatients (n = 119) at a residential drug treatment center. The variable was which of four different psychotropic medications had been prescribed for each patient. Each medication had been prescribed in roughly the same proportion (to about three-fifths of the patients, p > .20). So we cannot reject the null hypothesis. We must declare that there appears to be no pattern of one medication being favored by the psychiatrists who do the prescription at this facility.

The biggest problem with the sample vs. norms design is all the confounding variables that come into play with how the sample was selected. Suppose I have national poll (for my norm) showing that 48% of U.S. voters supported Hillary Clinton for President. My sample of community college students (n = 84) from California's Inland Empire shows that 63% would support that candidate (p < .05). So I reject the null hypothesis. But should I conclude that *students* really like Hillary? Or is it that she is more popular with *younger* voters? Or with *Hispanic* voters? Or maybe because most of the students are *female*? Or is it something geographical about *California*? These factors were not controlled in my sampling, and therefore the sample differed from the population in several key respects.

Correlational

A slightly better approach is *correlational*: we measure (at least) two variables and correlate them. Let's take my Inland Empire community college student sample (n = 84) and let's measure both the variable of gender (as an independent, predictor variable) and the variable of support for Hillary Clinton (as the dependent, criterion variable). I find a correlation of +.34 (p < .01). We can reject the null hypothesis. The women in this sample were more likely than the men to support Hillary. We have effectively controlled the variables of geography (they were all from the Inland Empire) and student status, but it is possible that some other confounding variables might be lurking. Could the females might be more likely to be poor single parents (and therefore more predisposed to Democrats)? If we had also measured marital status, parental status, and social class, then we could account for the strength of those correlations as well. Otherwise, we might have only spurious correlations leading us into post hoc fallacies about causation.

Separate Groups

Separate groups designs are no different from correlational designs if the grouping was just another measured variable (as it was in the above example of gender). The only time that separate groups designs have a real advantage over the correlational is when the grouping is randomly assigned, and the treatments are manipulated, thus giving us an experiment that can resist most confounding variables.

Repeated Measures

Repeated measures designs avoid the aforementioned problems of the other designs, but the drawback is that we have to get more than one measure on the dependent variable from each subject. Here are some examples of repeated measures designs.

- Did the workers (subjects) increase productivity (dependent variable) after training (independent variable)? This would be a within-subjects experiment, comparing the before training period (first measure) to the after training period (second measure).

- Do voters (subjects) have a higher favorability rating (dependent variable) of Hillary Clinton (first measure) or Donald Trump (second measure)?

- Do adults (subjects) become more religious (dependent variable) as they age? We could measure religious attendance at age 20 (first measure) and again at age 60 (second measure) for these same persons as they get older.

- Do married couples (the sampling unit) show that husbands (first measure) report higher marital satisfaction (dependent variable) than their wives report (second measure of the dependent variable)? If we just had a sample of individual men and women, and did not know who was married to whom, that would be a separate groups design: male vs. female. However, when we can look within each couple and compare a specific husband’s scores to those of his particular wife, that is a repeated measures design, giving us more statistical power because it controls for the inter-subject variation on background variables.

Unfortunately, each of these repeated measures designs opens us up to many new confounding variables that may be producing differences between the measures. Anything that changes over time, in addition to the variables being measured or manipulated, can impact the results. In a before and after treatment study, we have to consider the natural course of the disorder being treated. Some (like the common cold and depression) tend to improve regardless of treatment, while other disorders (e.g., dementia) tend to progressively deteriorate, regardless of the treatment. Some forms of performance improve due to practice or even just familiarity with the test, while other measures of performance deteriorate due to *boredom* or *fatigue*. Another major factor is *attrition* (loss of subjects from the sample), especially if the subjects more likely to score higher (or lower) on the dependent variable are less likely to show up for the second measuring period. Consider again those examples of repeated measures.

- Did the workers' initial performance measures take place soon after beginning their jobs, when they were not yet familiar with their tasks? Could the subsequent improvement be due to the natural course of improving over time with on-the-job experience rather than the specific training that was introduced?

- Was the order of how the candidates were presented a factor that was controlled, randomized, manipulated or even measured? Maybe presenting Trump first reminded voters that Hillary had all those scandals. Maybe presenting Hillary first reminded voters that Trump was insulting to women.

- By the time we get to age 60, do we even have a representative sample of the *cohort* that started out in 1957? Perhaps the least religious members of that cohort lived more on the wild side and did not survive to the age of 60.

- Do men and women experience marital satisfaction at the same stages of a marriage? Perhaps these couples were drawn largely from the early years of marriage where the men are more satisfied, but if it had been more elderly couples, the women would have reported more satisfaction.

So, no one design is always both easier AND better. Each design is a trade-off of convenience vs. dealing with confounding variables. Your proposal won't be perfect, but try to deal with confounding variables as best as you can within what is possible for you. The video summarizes the four research designs.

The Proposal

A formal proposal for psychological research, in the most bare-bones format, would have to answer five questions.

WHAT IS THE TOPIC? If it is psychology then it can be stated as a question involving some aspect of behavior, personality, attitudes, mood, performance, or choice. In other words, the topic is defined by the dependent variable.

WHO ARE THE PARTICIPANTS? Describe your sample of convenience, including an estimate of its size. From whom will you get data? (e.g., 50 students coming out of the library, two dozen participants in a Bible study, 20 lab rats).

HOW WILL DATA BE COLLECTED? Most student data will probably come from a questionnaire. Other possibilities are field counts, traces, and archives. Qualitative data can come from interviews (e.g., focus groups), participant observation, or analysis of visual or textual data. Be very precise about your operational definition of each major variable.

WHICH DESIGN WILL BE USED? Hypotheses must be tested by one of four designs: 1) comparison of entire sample to pre-established norms (e.g., national polls, census data); 2) comparison of separate groups (e.g., men vs. women, experimental vs. control); 3) comparison of repeated measures (e.g., attitudes about different aspects, before and after); 4) correlations between variables.

WHAT ARE YOUR HYPOTHESES? State at least one hypothesis, or several hypotheses: predictions(s) of correlation(s) or difference(s) you expect to find, especially those that might be consistent with some theory with which you are familiar.

Usually, you can complete a proposal quickly if you it in this order. First, determine which population you have quick access to. (You can't do studies of the schizophrenic population if you don't have access to a psychiatric hospital). Second, determine which measures you have access to and are appropriate for your sample. This will determine your collection of data. Then choose a design appropriate for your sample and your data measurement. Give those measurements and that design, what hypothesis could be tests? Now your last point is formulating the topic: just what did you end up deciding to study.

This video gives a good visual and motor mnemonic for remembering this order.

Perhaps what is most important is that you do your proposal ***quickly!!*** It does not have to be perfect; it just has to be good enough to get accepted. The most important thing about a good enough proposal is that it is done quickly. If it is not good enough to get accepted, your instructor will tell you and give you feedback on how to make it acceptable (as long as the deadline has not passed). So, get busy on that proposal ***now!!***

Chapter #5: Measurement & Coding

"Knowing what to measure, and how to measure it, can make a complicated world less so."

-- Levitt & Dubner (2014)

While some independent variables can be manipulated, controlled, or randomized, these procedures cannot be used on dependent variables. The researcher cannot randomize, intentionally vary (manipulate), or intentionally control (i.e., set to a constant level) a dependent variable. By their very nature, dependent variables depend upon the subject's choice or performance. It is possible that a given dependent variable will (in a given sample or population) be constant if each subject ends up having the exact same score. However, if these are the results, then we cannot do any hypothesis testing. In order to do test a hypothesis, the dependent variable defining our topic must vary and be measurable.

Psychometrics is the branch of psychology concerned with tests and measurements. All variables (and constants) must be clearly defined conceptually and operationally. Do not try to define a variable by looking in a regular dictionary (or Wikipedia). Use a specialized dictionary (or encyclopedia) such as

APA Dictionary of Psychology
Washington, DC: American Psychological Association
BF31 .A65 2007 University of Redlands Library
Crafton Hills College Library

International Dictionary of Psychology
New York: Crossroad

The *operational definition* of a variable is how it is to be actually measured in practice. This requires a decision as to what kind of scaling we will use: ratio, interval, ordinal, or nominal.

Ratio/interval Scaling

The most precise way to measure a variable is on a ratio (or interval) scale. This means that the variable is represented by a number or score, such that a higher score represents being high on the variable and a lower score represents being low on the variable. There is a distinction between *ratio* scales (which have a true zero point) whereas an *interval* scale (e.g., IQ score, Fahrenheit, Celsius) does not. A true zero point means that a subject who scores a zero has none of the variable. True zero points apply to measures of time, length, area, volume, incidents, customers, income or units of production. However, to say that it is zero degrees outside does not mean that there is no temperature or heat. It is just an arbitrary point on a thermometer.

Furthermore, a true ratio scale has proportionality, such that a participant who scores twice as high on a test has twice the level of the variable. This does not hold for interval scales. A person with IQ of 120 is not twice as smart at one with 60. When the thermometer says 80 degrees, it is not twice as hot as a day with 40 degrees.

The most precise ratio scales are *continuous* ratio, which means that they can be divided into decimals and fractions. Measures of distance, volume, and time are continuous ratio scales. Variables that have indivisible units (e.g., units produced, accidents, customers) are *discrete* ratio scales.

The good news is that you do not have to worry if a scale is continuous or discrete (or even ratio or interval). If each participant is scored on a variable by receiving a number, that means that you can use *parametric* statistics (e.g., mean, standard deviation, Pearson correlation, t test, ANOVA), assuming that the scores on that variable, in the population, approximate a normal *Gaussian* curve (a.k.a., bell curve).

Ordinal Scaling

A common way of measuring dependent variables, especially attitudes, is the use of *ordinal* scaling, which is less precise than interval or ratio. Ordinal scaling means that although each participant does not receive a number to measure a variable, the participants can be ranked (compared to each other), such that we know who is first (highest) on that variable, and then who is next, and so on until we get to the last subject (who is lowest on that variable). Ordinal scales allow for ties on rankings, and in many cases we have so many ties that we might just as well speak of different levels of a variable.

For example, many attitudes are measured on a five-level *Likert* scale: completely agree, mostly agree, not sure, mostly disagree, completely disagree. Those who answered that they "mostly agree" have reported *more* agreement (a *higher level* of agreement) than those who are "not sure" but *less* agreement than those who "completely agree." Many Likert scales (like the ones in the Texas Ten Item Personality Inventory measuring the Big Five personality traits) use seven levels of agreement, putting in "slightly agree" and "slightly disagree" on either side of "not sure."

We could make a Likert Scale with an even number of levels by eliminating the “not sure” in the middle.

Another example of an ordinal scale would be a frequency scale. The question might be “How often do you go out to the movies”? The response format could be: at least once a week, several times a month, several times a year, rarely. Someone who answers “several times a month” goes more frequently (*more* often) than someone who goes “several times a year” but not as frequently (*less* often) as someone who goes “at least once a week.”

Another ordinal scale would be evaluational: “Rate your boss's performance” and the response format might be excellent, good, fair or poor. Someone who said “good” would be rating his boss better than someone who only said “fair” but not as good as someone who said “excellent.”

Other examples of ordinal scaling would be increasing/ decreasing levels of certainty (e.g., definitely / probably / possibly / no way) or intensity (e.g./ extremely / very / somewhat / slightly / not at all). A self-rating against an average would also qualify as ordinal: very much above average, slightly above average, about average, slightly below average, very much below average. Another way to phrase this might be to use numbers as sort of benchmarks, such as, “Rate yourself, compared to others your age in terms of commitment to serving your community” and the response format would be: highest 10%, … lowest 10%. We could have several levels in between. Another example of ordinal scaling is where interval or ratio scales have been collapsed to levels or ranges, such as this one for age: under 20, 20-29, 30-39, 40-49, 50+. Whenever the top number (or bottom number) is given as a range of scores rather than a specific number, we have ordinal scaling, rather than interval or ratio.

There is disagreement among statisticians whether certain scales are more interval or ordinal. For example, take Cantril's 0 - 10 ladder of life satisfaction, frequently used by the Gallup organization and Mitofsky. Here is an example of the former's use of the Cantril scale to measure well-being. Although the subject is responding to the question by indicating a specific number between zero and ten, it is not clear that there is any way to objectively demonstrate that there are equal intervals of some real-world external entity to go along with these numbers. These numerical ratings might just be considered eleven different levels that the subject may select based upon a self-report.

Another scale where it is unclear whether it is really interval or ordinal would be a "feeling thermometer" where the subject can select a number between 0 and 100, where 0 = not at all, 25 = mildly, 50 = somewhat, 75 = very, and 100 = extremely. One problem with this scale is that these example numbers become anchored in the subject's mind, and tend to be selected more often than other numbers in between (e.g., 23, 37).

Ordinal scaling qualifies for the use of statistics such as the median, percents, and many sophisticated *nonparametric* inferential tests such as Friedman, Wilcoxon, Kruskal-Wallis, Mann-Whitney and Kolmogorov-Smirnov (all of which will receive further explanation in the next chapter). Some statisticians argue that these aforementioned ordinal scales could use parametric statistics if the variable(s) in question have a *normal* (Gaussian) distribution, otherwise nonparametrics would be called for, especially when the sample size is small.

Nominal Scaling

The least precise quantitative measure of variables is to use *nominal* scaling (*categorical*). If there are only two categories into which every participant can be categorized, then we have *binary nominal scaling* (also known as *dichotomous*). Examples would be: male/female, pass/fail, experimental/control. Whenever you impose a yes/no question, you end up with binary nominal scaling. Make sure that your two categories are really mutually exclusive, and do not permit a both/and (or neither/nor) rather than either/or response. For example, "Are you talking chem or bio this semester?" could be answered chem, bio, both, or neither. It would be better to see this as two questions: "Are you taking chem?" and "Are you taking bio?" Binary nominal scaling requires the use of percents as descriptives and nonparametric inferential statistics (e.g., Chi Squared).

Adding more categories to nominal scaling gives us multiple nominal variables. Examples would be major (e.g., business, psychology, engineering, etc.), political affiliation (e.g., Democrat, Republican, Green, Independent, etc.), or brand of automobile owned (e.g., Toyota, Chevy, Ford, etc.). Remember, in order for one variable measured on a multiple nominal scale to be appropriate, these categories must be mutually exclusive (at the present time I own a Chevy, Chrysler, Nissan, Mazda, Hyundai and several Fords). Another problem is that we end up with a lot of categories with only a small number of subjects in them (e.g., people who belong to minor political parties) and that creates problems for the inferential statistics.

An alternative is to create several binary nominal variables: Do you have a Chevy: yes/no? Do you have a Toyota: yes/no? Even if the question was asked in a multiple nominal format, it may be better, statistically, to switch to binary nominal coding for these variables: have several variables (i.e., columns in Excel), one for each alternative as a yes/no variable.

Since any measurement can be reduced to a less precise measurement, ratio, interval and ordinal scales can also be reduced to the yes/no of binary nominal. For example, we could ask the subject's age (How old are you?) and get a ratio discrete answer. For example, my mother is 89 years old. We could collapse answers recorded on that scale to an ordinal scale: Which age level does the subject fit on?

Under 20	20-29	30-39	40-49	50-59	60+

Notice that this involves a loss of precision. Indeed, it would even throw myself, my wife and my mother into the same age category and ignore the precise differences in how old we are. We could even reduce ordinal and ratio scales to binary nominal: Are you over age 40? Yes/No. Think of this as an ordinal scale where the number of levels have been reduced to just two: over 40 and under 40. It is even less precise because now my wife and most of my nieces fall into the same category.

So you can always go from a more precise scale to a less precise one: ratio to ordinal, ordinal to nominal. However, you cannot go in the other direction. If your original questionnaire measured age as "Are you over 40"? then we would not know who was over 70 and who was a mere 56. So, the guideline is, collect your data on as precise a scale as possible. You can always simplify to categories and percents later.

More on scaling can be seen in this video.

Validity & Reliability

Good psychological measures should be reliable and valid as well as precise and practical. (Do not use regular dictionaries to understand validity or reliability.) A *reliable* psychological measurement is one that has consistency of measurement: the same subject gets a similar score on the same variable. There are four forms of reliability.

Form of reliability	Description / Example
Test-retest	When a subject takes the test again, he gets a similar score.
Internal	Subjects who pass the first item on the test are more likely (than other subjects) to pass the next item on the test.
Inter-rater	If we need raters (judges) to score subjects on a variable, the subject will get a similar rating, regardless of who is performing the evaluation.
Alternate form	Whether taking the computerized version or the face-to-face version of the test, the subject gets a similar score

When we say that each subject "gets a similar" score we are comparing the subject to his/her own performance (at a different time, on a different version of the test, on a different part of the test, or when evaluated by a different examiner). We are not saying that a reliable test means that each subject scores the same as every other subject in the sample. If that happens, then the variable has become a constant, and we cannot establish reliability or test any other hypotheses.

Reliability is established by correlational research: the same subjects who score high on the test (the first time it is given) should be same ones who score high on that same test (the second time it is given). Correspondingly, the same subjects who score low on the test (the first time it is given) should be the ones who score low (the second time it is given). Reliability is represented by a correlation coefficient. Specialized coefficients (e.g., *Cronbach*, *Kuder-Richardson*) are used to measure internal reliability. More on reliability can be seen in this video.

A *valid* measure of a variable measures the variable that is supposed to be measured (as opposed to some other variable that may be easier to measure). We start with the *face* validity of a measure: does it look like it measures the right variable? We move on to *construct* validity: does it measure the entirety of the variable? We go on to *criterion* validity: does the new measure correlate with variables known to correlate with the target variable? We then consider *discriminant* validity: does our measure avoid contamination from other variables not really related to the target variable?

Validity	What this involves / Example
Face	Does the item look like it measures the right variable, or does it look like it is really measuring something else? **Example of Bad Test:** **Is Jones depressed? Jones is angry with his wife, therefore he is depressed. But, anger is a different variable. Anger does not have face validity as a measure of depression.**
Construct	Is the entirety of the variable measured by the test, or is only a limited part of the variable included? **Example of Bad Test: Jones has had insomnia for three weeks, therefore he is depressed. Insomnia is only one symptom of depression. Other symptoms need to be included.**
Criterion	Does the test correlate with other variables known to correlate with the target variable? **Example of Bad Test:** **Jones scored high on a test of life satisfaction. If Jones were really depressed, we would expect him to score low on a test of life satisfaction.**
Discriminant	Does the test avoid contamination with other variables? **Example of Bad Test: Jones has had numerous physical complaints over the past three weeks. These could be symptomatic of depression, but they could also be due to other factors, such as a real physical illness.**

This video explains the basic concept of validity, while this one uses an analogy of the size of an egg, and this one looks at the depth of a well.

Obviously, establishing the reliability and validity of a test, scale or simple measure of a variable requires well-planned research beyond the scope of this course. Therefore, don't assume that your research will be sufficient to establish the validity and reliability of a measure. It is better for you to use a test that has already been validated (e.g., Rosenberg Self-Esteem Scale, Texas Ten Item Personality Inventory, Center for Epidemiological Studies Depression Scale, Geriatric Depression Scale, Beck Depression Inventory). Such a test usually has norms that your entire sample can be compared with. If you cannot find an established scale (or poll item) to measure the variable you are seeking to measure, then come up with the best one item you can, using an ordinally scaled response format of four or five levels. Do not come up with a ten item scale where you just add up the points from each and assume that the total represents a valid and reliable measure of the variable.

A *composite* variable is where we take two (or more) different variables and link them together with a formula to derive a third variable. Composite variables can be used as criterion variables (e.g., measures of overall performance) or as predictors (e.g., measures of overall stress). One example would be in combined Olympic events, such as the decathlon. Each athlete gets so many points based upon each of the ten events, and the athlete with the highest overall score gets the gold medal. Another example would be typing proficiency, usually measured by words per minute minus number of errors. As with multi-item scales, we should not merely assume the validity and reliability of composite variables. Indeed, separate studies should be done prior to their use in order to verify these qualities of the composite variable.

A good thing to keep in mind when coding the data for a composite variable (or multi-item scale) is to code a column for each specific item or component. This will enable us, should we choose to do so at a later point in time, to go ahead and do an *item-analysis*. In this way, we may find out that our experimental intervention (e.g., training with a new typing program) had little impact on overall typing proficiency, but it was pretty good in reducing error rate (though not overall speed). Or, we might find out that weight training helped decathletes improve their performance overall, but actually hurt it in a specific event.

Where to Find Tests

Here are some good sources of established psychological tests.

Measures for Clinical Practice & Research
Fischer, Joel & Corcoran, Kevin
New York: Oxford University Press
BF176 .C66 2007 Crafton Hills College library

Tests in print: an index to tests, rest reviews, and literature on specific tests.
Highland Park, NJ: Gryphon Press.
Z5814.E9 T47 University of Redlands library

Mental Measurements Yearbook.
Highland Park, NJ. Gryphon Press.
Z5814.P8 B932 University of Redlands Library

Directory of Unpublished Mental Measures.
Washington, DC: American Psychological Association.
BF431 .G625 University of Redlands library

An alternative for attitudinal measures is to use poll questions developed by major polling organizations. The clarity of these questions (and their response formats) have been field tested. Furthermore, there are usually national norms on these questions that your sample can be compared with.

Polling data (and some background data) can be found at the General Social Survey maintained by the National Opinion Research Center of the University of Chicago. The annual study of entering freshman is conducted by the Higher Education Research Institute at UCLA. The best private polling organization in the U.S. is Gallup
Click on "topics" and you will see quite an array dealing with political, economic, and personal questions. There are other polling organizations in the U.S., such as Roper and People-Press. Zogby is one of the best about the Middle East. Barna specializes with Evangelicals, but they also have studies about Millennials in the workplace. Polling Report summarizes many other polling organizations. In Mexico, the best polling is done by Mitofsky. Another place to find measures of variables that have already been validated (or at least passed a field test) would be with previous studies done on the same variable.

If you are not using a question (and response format) that has already had its reliability and validity established, it may be necessary to conduct a limit trial run known as a *pilot* study in which a small number of people (representing the population you wish to study, but not part of the actual sample you will be running your statistics on) are given the question just to see if they understand what is being asked and to see if their range of responses avoids both *ceiling* effect (i.e., most people picking the highest level answer) as well as *floor* effect (i.e., most people picking the lowest level answer).

Coding

One important factor to keep in mind as you are selecting the operational definitions for your variables is how you will end up coding them. Quantitative data will end up in a spreadsheet program. (Think of a spreadsheet program, such as Microsoft *Excel* or *Google Sheets*.) Use each column for a different variable and each row for a different participant (one participant per row, one row per participant). If you cannot conceive of how your data will go into that kind of rows and columns configuration, STOP right now. You need to rethink how you will measure those variables (or if those are the variables that you can measure, or whether you should be using qualitative measures at this time).

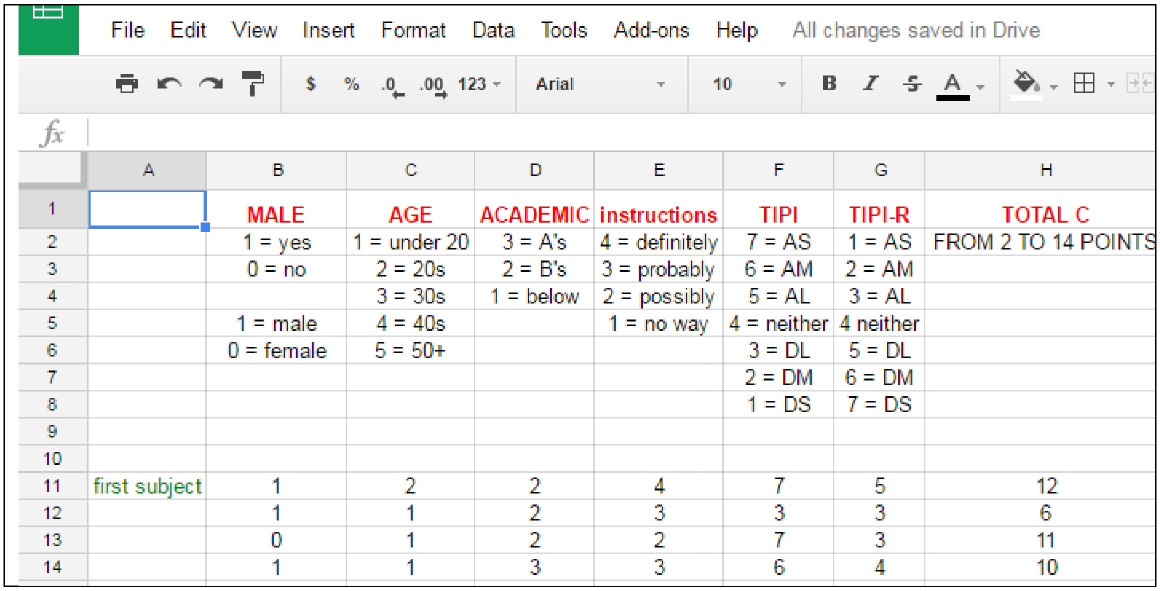
What might be a wise move to preserve clarity and transparency would be to use the top ten rows just to clarify each variable. Put the names of the variables in the first row (starting in B1). Name each variable in such a way so that it is clear what a high score means. For example, AGE is a good name for a variable, because we know that a higher score means older and a lower score means younger. If we see DEPRESSION we are going to assume that a higher score means more depressed and a lower score means less depressed. But if what you are doing is actually measuring life satisfaction (perhaps on the Cantril ladder) and then inferring the subjects' level of depression, it would be better to call the column something like LADDER or LIFESAT. If you are going to call it DEPRESSION, reverse score it. Any measure of performance should be coded such that a high score implies better performance and a low score implies worse. This is easy if we are just looking at a point score on a test. In baseball this works with variables like a batter's runs scored or a pitcher's strikeouts. It also works with composite variables like batting average and slugging. However, in some sports, low numbers mean better performance: a pitcher's earned run average, a golfer's strokes, a runner’s time. In these cases, clearly label the columns as ERA, STROKES, TIME in order to avoid future confusion that you, your research colleagues, or someone else looking at your charts might have. When performance is measured on a binary nominal scale (e.g., pass/fail) do not score these as one point for pass and two for fail: use pass = 1, fail = 0. When you have a yes/no variable: yes = 1, no = 0. When you have a separate groups experiment: experimental = 1, control = 0.

If you have a nominally scaled variable where it is not obvious which condition is higher or lower, try to clarify that in the very name of the variable you put at the top of the column in row 1. For example, if you are using numeric coding, don’t call a variable gender, sexual orientation, religion, or school attended. Instead, use these column headings: MALE, LGBT, CATH, CHC. Now, each of these can be yes/no scored (numerically with ones and zeros) and we know that someone who scores high on each variable is a Catholic gay man who attends Crafton. Otherwise, how will we interpret a correlation between any of these variables?

Use the next few rows to explain your scoring. This is just another way to clarify some of the things suggested in the previous paragraphs. Use rows two through ten to explain what is a 1 and what is a 0. This is especially helpful when you have ordinally scaled responses and you need to put those into numerical coding. So, you use these rows to explain that strongly agree = 5, mostly agree = 4, don’t know = 3, mostly disagree = 2 and strongly disagree = 1; or that daily = 5, weekly = 4, monthly = 3, rarely = 2, never = 1. Or perhaps you want to take everything down a point so that your lowest answer is a zero. You just need to be consistent in your scoring, and transparent so that anyone else looking at your data understands what you did, how and why.

We can also achieve more clarity and transparency by using the left column (column A) to have notations about which participant or group we are referring to. For example, if rows 11 through 27 refer to the 17 subjects in the experimental group, and rows 28 through 39 may refer to the 12 subjects in the control group, then column A might contain some labels to reiterate that assignment.

Here is a screenshot of a Google Sheet. The criterion variable was how well the subject estimated his/her own ability to follow instructions. The predictor variables were gender, age, academic performance and the conscientiousness scale on the Texas Ten Item Personality Inventory (with both component questions shown here, notice that the TIPI-R was reverse scored).



	A	B	C	D	E	F	G	H
1		MALE	AGE	ACADEMIC	instructions	TIPI	TIPI-R	TOTAL C
2		1 = yes	1 = under 20	3 = A's	4 = definitely	7 = AS	1 = AS	FROM 2 TO 14 POINTS
3		0 = no	2 = 20s	2 = B's	3 = probably	6 = AM	2 = AM	
4			3 = 30s	1 = below	2 = possibly	5 = AL	3 = AL	
5		1 = male	4 = 40s		1 = no way	4 = neither	4 neither	
6		0 = female	5 = 50+			3 = DL	5 = DL	
7						2 = DM	6 = DM	
8						1 = DS	7 = DS	
9								
10								
11	first subject	1	2	2	4	7	5	12
12		1	1	2	3	3	3	6
13		0	1	2	2	7	3	11
14		1	1	3	3	6	4	10

Let’s go down to row 11 for the start of our first subject (which means if our sample size is 50, our last subject’s data will be in row #60).

The number entered into each cell (starting with B11) is the value (e.g., score) that particular subject obtained on that particular variable. The above screenshot shows that our first subject (row #11) was a male, in his 20s, whose grades were mostly in the B range. He rated himself as "definitely" one who can follow instructions, and gets a combined score of 12 on the TIPI conscientiousness scale. Our next subject (row #12) was also a male, but under age 20, also a B student. He rated himself as "probably" able to follow instructions, and only had a combined score of 6 on the conscientiousness scale.

For interval and ratio scales, we would just enter the number (score) of that subject on that variable, as indicated in this video. For ordinal scales involving ordered levels, convert to scores so that the highest level gets the highest score (e.g., excellent gets a 4 and poor gets a 1; completely agree gets a 5 and completely disagree gets a 1; the most frequent response gets a 4 and the least frequent gets a 1). When asking if a particular trait fits an individual, the response pattern might use this coding: definitely true of me = 5, tends to be true of me = 4, unsure = 3, tends not to be true of me = 2, definitely not true of me = 1. This video gives more examples of coding for ordinal scales.

An exception to using numbers for ordinal levels comes if you know you will use these levels to form separate groups (and perhaps then run an ANOVA or Kruskal-Wallis on the criterion variable). If this is the case, you can use an abbreviation instead of a number in each coding, such as EXC, GOOD, FAIR, POOR or DEF, PROB, POSS, NO.

If you have ranks, instead of scores or levels, then you will have to *reverse* score. If you had a sample of twenty-five subjects, ranked first (highest) through twenty-fifth (last), you would have to give 25 points to first place, 24 points to second place, etc. Never violate the basic principle that a participant who scores higher on a variable must get a higher score and a participant who scores lower on a variable must get a lower score.

For binary nominal variables, use "*dummy*" coding such that the category highest on the variable gets a 1 and the other category gets a 0 (e.g., yes gets a 1 and no gets a 0, pass gets a 1 and fail gets a 0; over a certain age get a 1 and under that age gets a 0). This video demonstrates dummy coding.

When you have a binary nominal scaling of a variable that does not have a naturally high level or low level, redefine the variable as yes/no. For example, if we take gender, we cannot say that male or female is higher on that variable. So, for clarity, redefine the variable as a yes/no question: Is the participant male? Don't refer to the variable as gender anymore, but call it "male" and score it as yes (male) = 1 and no (female) = 0. That way if you see a negative correlation between that variable and another variable (e.g., academic performance) you will be able to properly interpret it: women (non-males) had higher academic performance, men had lower academic performance.

The worst kind of scale to have to code is *multiple nominal*. Suppose the variable is religious affiliation (i.e., denomination). Suppose you have ten categories: Roman Catholic, Latter-day Saint, Seventh-day Adventist, Jehovah's Witness, other Christian, Jewish, Muslim, Buddhist, other religion, and no religion. Does it make sense to code these ten categories 1 through 10? If the scale is really ordinal (perhaps you are measuring the difference between a denomination's doctrine and that of the Catholic tradition, so it would make sense to give "no religion" the highest score and Catholics the lowest, but then you would have to justify saying that LDS are closer to Catholics than SDA are or that Jews are closer than Muslims.

If you cannot make such ordinal assumptions, some strange numerical operations will get performed. When you do averages, you are assuming that the average between a

Catholic and a Seventh-day Adventist is a Mormon, and the average between a Buddhist and a Jew is a Muslim. If that doesn't make any sense, neither will the correlations that come out of this scoring.

There are two solutions. One is to keep multiple nominal scales clearly labeled as nominal. Don't use numbers, just use abbreviations (Group H, Group I, Group J, etc.). That means you won't be able to do a correlation matrix, but will have to use Chi Squared or ANOVA for your inferential statistics.

The other solution is to convert a multiple nominally scaled variable into several variables, each with binary nominal scoring. In other words, make each category a separate binary nominal variable (yes/no) and then dummy code.

So, have one column for Catholics, one for LDS, one for SDA, one for JW, etc. Each subject will get a score of 1 in one of the columns (the one representing his/her denomination) and a 0 in all the other columns. A subject who is Mormon would have the ten religion columns look like this:

RC	LDS	SDA	JW	oC	Jew	Mus	Bud	other	none
0	1	0	0	0	0	0	0	0	0

This has approach has several advantages. One is that when you do a count, you will get the number in each category of the entire sample (14 Catholics, 4 LDS, 5 SDA, 1 JW, 20 other Christian, 1 Jew, 2 Muslims, 0 Buddhists, 1 other religion, and 4 no religion). When you do a "mean" for each column, you will get the percent (expressed as a decimal) of the entire sample in that category. Another useful feature is an additional opportunity for error check. Add up the count from each column devoted to these religious variables and it should equal the sample size. Add up the means from each column and it should come close to 1.00 (i.e., 100%). If you cannot explain your divergence from this number as rounding errors, you made a mistake in data entry.

An exception for using dummy coding for nominally scaled variables comes if you will only use them as grouping variables in a separate groups design, in which case you can code them as EX and CN for experimental and control, or MALE and FEM for gender. However, if you like the idea of creating a great correlation matrix for all variables, use numbers to code all variables, including all the ordinal and nominal ones.

Missing Data

Perhaps the greatest problem you will face in coding is what to do about missing data: e.g., the subject in row 26 did not answer the question about age, so you have nothing to enter into cell C26. Let’s review potential solutions (ranging from the completely unacceptable to the more tolerable). The best solution, under most circumstances, is to eliminate that subject (and that means eliminating the entire row: all the data on all the variables for that person).

Trying to figure out what answer the subject would have given is known as *imputation*. One common approach is to give the sample’s average value on that variable. Another approach is to input a value from another member of the sample, selected at random. Imputation has less adverse impact if the sample size is large, the number of missing data cases are small, and there is no underlying pattern of who the missing data cases are. Since it is rare that all three of these assumptions hold (at least with my data) I do not use imputation.

Action	Problem	Acceptability
Enter a 0 value	This will grossly distort the average and any correlations involving this variable.	Worst solution, never acceptable
Enter the average value of the sample or group (or a value from another subject, randomly selected)	This assumes that the subject not answering this question is more likely to be similar to the subjects who did answer.	Bad solution
Enter the average value of that subject on similar measures	This can be done where the missing item is a part of a composite variable, such as one question on a test. This assumes that a subject who passed the other items would have passed this one, or would hold a similar attitude.	Less than optimal solution
Eliminate the variable (which means eliminating all the data from the other subjects who did answer the question).	This can be done when you have many variables, and the variable on which you have missing data is not central to the hypotheses being investigated. You cannot do this if the missing data comes on an item used in constructing a large scale.	Tolerable under some circumstances, but not if the question is part of a scale measuring the criterion variable.
Eliminate the subject (which means eliminating all the data from the other variables on which the subject gave a scorable answer).	This can be done unless your sample size is extremely small to begin with. (A better alternative might be to try a new sampling.)	Best practice.

Elimination of the subject (or variable) with the missing data is usually the best solution. If your data analysis is primarily a correlation matrix with all variables, eliminate all of that subject’s data. (That is what most statistical problems do as a default when performing a correlation matrix, automatically reducing the sample size). Alternatively, if you are using different statistical techniques for your different hypotheses, you could simply remove that missing data case from only those calculations involving the missing data. For example, if in the previous example about the ability to follow instructions (the criterion variable) assume that in all fifty cases, there were data about conscientiousness and gender, but one subject did not answer the question about age. We could still use that subject in correlating gender to the criterion variable, and conscientiousness to the criterion variable, but when we got to the correlation with age, that subject would be removed.

The best approach is to prevent or reduce missing data situations by having a questionnaire that is short, clear, and easy to answer.

A similar problem arises when you have reason to believe that a given subject did not answer the questions seriously (e.g., someone just circled all the answers on the right side, regardless of the question): just eliminate that subject instead of trying to figure out how that anonymous person would have answered he/she had taken the time to give real descriptive answers.

The Future of Technology

If there is an Achilles Heel to most measurements in psychology (especially for criterion variables) it is an over-reliance on subjects' self-report. People over-estimate how kind and moral they are, and even their intelligence (often attributing a lack of academic performance to external obstacles).

Newly developing mobile and biometric technology involves the possibility of data gathering techniques that have many advantages over paper and pencil self-ratings. The *big data* captured on a smart phone or watch are more ecologically valid, in that they are measured in real time and not distorted by a subject's memory (or motivation or self-presentation). These digital data can have dense sampling, collecting measurements over microsecond time periods, and bring a precision incomparable to that of a subjective rating on a five level ordinal scale. More data can be obtained from fitbits, RFID chips, GPS, cameras, motion sensors, Point of View eye tracking, brain scans, even telemetric data about patients starting from initial contact at the scene, through the ambulance ride, and then in hospital.

The great benefit of the *analytics* permitted on these big data is not just that the sample size is larger, and more variables have more frequent measures, but the data are automatically, passively, and directly collected, and not distorted by the subject's need to present self in the best possible light or requiring inter-rater reliability. Furthermore, the sample is more inclusive (and therefore more representative) because it is harder to opt out. The sampling is ongoing, not stopping at a certain n or date. The data analysis can be done immediately, and then specifiable to a particular sub sample.

References:

Levitt, Steven D. & Dubner, Stephen J. (2014)
Think Like a Freak. New York: William Morrow.

Chapter #6: Statistics

Descriptive Statistics for a Variable

Descriptive statistics describe a variable (or the relationship between variables) and include measures of central tendency (average), dispersion and correlation.

Rounding off is an important part in reporting our statistics. The designated digit is the place that we want to round to (e.g., hundredths). Rounding down means leaving the designated digit as it is. Rounding up means raising the designated digit by one unit. To decide whether we should round up or round down, look at the number to the right of the designated digit. So, if we are supposed to round to the hundredths place, the number we look to in order to tell us whether to round up or down would be in the thousandths place. Here is a summary of the rules and four examples of rounding to the hundredths place.

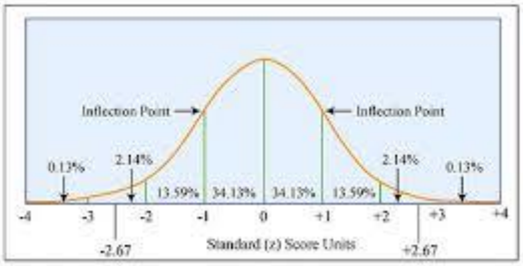
Digit to the right is	Then round	Example
0 or 1 or 2 or 3 or 4	Down	3.012 goes down to 3.01
6 or 7 or 8 or 9	Up	2.978 goes up to 2.98
5 (with something other than all 0's after it)	Up	6.0151 goes up to 6.02
5 (with nothing after it)	Up if designated digit is odd; but Down if designated digit is even	7.315 goes up to 7.32 4.065 goes down to 4.06

Generally, we *round off* to whole percents, and round off to hundredths place for correlation coefficients, means, standard deviations, t-scores and other inferential statistics. This video shows some basic rules for rounding off.

Percents are appropriate descriptive statistics for variables arranged in categories or (a few) ordered levels. Percents across the categories or levels of the variable demonstrate central tendency and dispersion. This video shows you how to do such a percent calculation.

When we have numerical data from interval or ratio scales, we can use measures of central tendency (sometimes called averages) such as the mode, median or mean. This video shows how to identify the mode. This video shows how to identify the median. This video shows calculation of a mean.

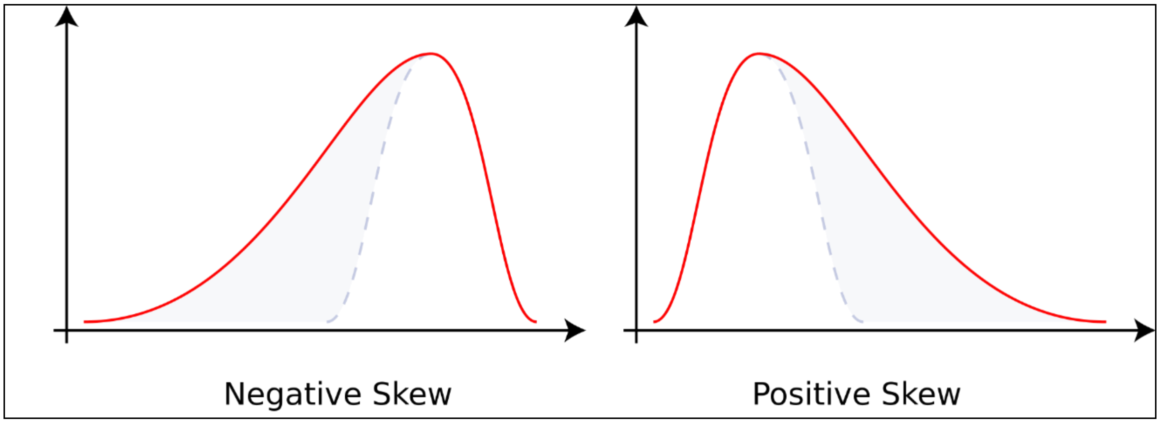
The *mean* is the most precise measure of central tendency, and is appropriate for *normally* distributed data sets (i.e., when the data are distributed according to the *bell-shaped* curve described by *Gauss*). That curve assumes that data have a single hump (mode) in the middle of the distribution of scores (but not necessarily in the mid-range of possible scores), and that both “tails” are symmetrical with a decreasing number of scores as we get further out from the mean. This video explains the normal distribution and other statistical curves.



Proportions of scores in specific areas under the normal curve. The inflection points are one standard deviation from the mean.

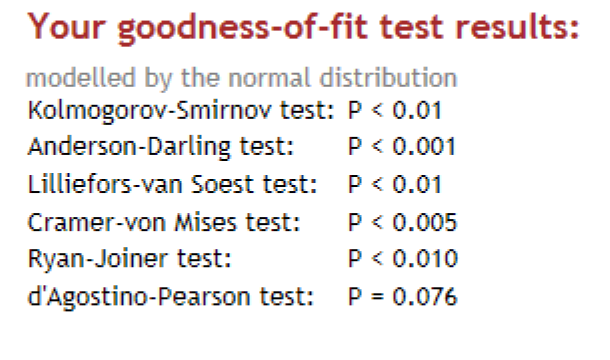
A normally distributed data set can use *parametric* statistics, while a lack of normality (especially with a small sample size) suggests the need for *nonparametrics*.

A normally distributed curve assumes that neither tail has been *truncated* (i.e., cut off by ceiling or floor effect) and that *outliers* have not produced a long tail on either the high end (*right, positive skew*) or low end (*left, negative skew)* of the range of actually observed scores.



You should not automatically assume that the distribution of scores in your sample is normally distributed for every variable. Indeed, most variables in student projects are not normally distributed. This lack of normality can cause major distortions in both descriptive and inferential statistics. You can just copy your data from an Excel (or Google sheet) column and paste it into this website to see if it is normally distributed. The statistical tests employed tell how much difference there is between your data set and one that is normally distributed. When p < .05, that difference is significant, and you may conclude that the variable is not normally distributed.

These p values show significant differences between this variable’s distribution and that of a normal distribution of scores (for every test except for d’Agostino-Pearson).



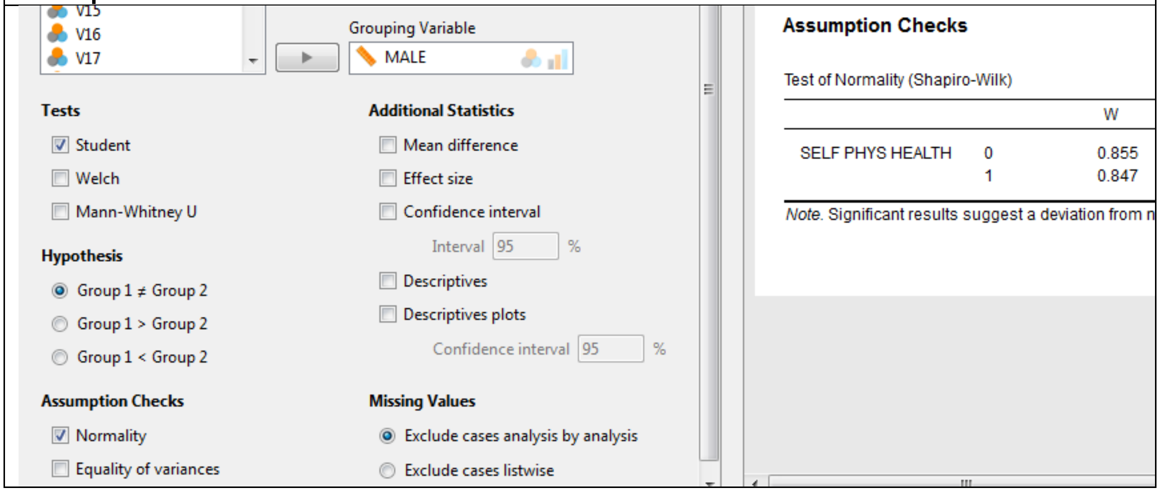
Your goodness-of-fit test results:

modelled by the normal distribution

Kolmogorov-Smirnov test:	P < 0.01
Anderson-Darling test:	P < 0.001
Lilliefors-van Soest test:	P < 0.01
Cramer-von Mises test:	P < 0.005
Ryan-Joiner test:	P < 0.010
d'Agostino-Pearson test:	P = 0.076

This site will run several “goodness of fit” tests at once. Most of these tests give similar verdicts, so I prefer to use two very different tests: the Cramer-Von Mises and the d’Agostino-Pearson. If the data pass both of these tests (with p > .05) I shall affirm that the data are normally distributed. Alternatively, if you are using the JASP program to perform a parametric inferential statistic (e.g., Analysis of Variance, t-test) one of the options is to test for normality using a Shapiro-Wilk test.

The p = .002 means that these scores differ significantly from the normal distribution, and therefore, nonparametrics are called for.



V15
V16
V17

Grouping Variable: MALE

Tests

☑ Student
□ Welch
□ Mann-Whitney U

Hypothesis

● Group 1 ≠ Group 2
○ Group 1 > Group 2
○ Group 1 < Group 2

Assumption Checks

☑ Normality
□ Equality of variances

Additional Statistics

□ Mean difference
□ Effect size
□ Confidence interval
Interval 95 %
□ Descriptives
□ Descriptives plots
Confidence interval 95 %

Missing Values

● Exclude cases analysis by analysis
○ Exclude cases listwise

Assumption Checks

Test of Normality (Shapiro-Wilk)

		W
SELF PHYS HEALTH	0	0.855
	1	0.847

Note. Significant results suggest a deviation from n

The *median* is a nonparametric measure and probably the most efficient estimator of central tendency for interval and ratio scales when there is a skew, because the median will not be affected by the distance of the most extreme scores from the center. The median is *robust*, and resists distortion. On the other hand, when the distribution follows the symmetrical Gaussian curve, the median equals the mean (and the mode).

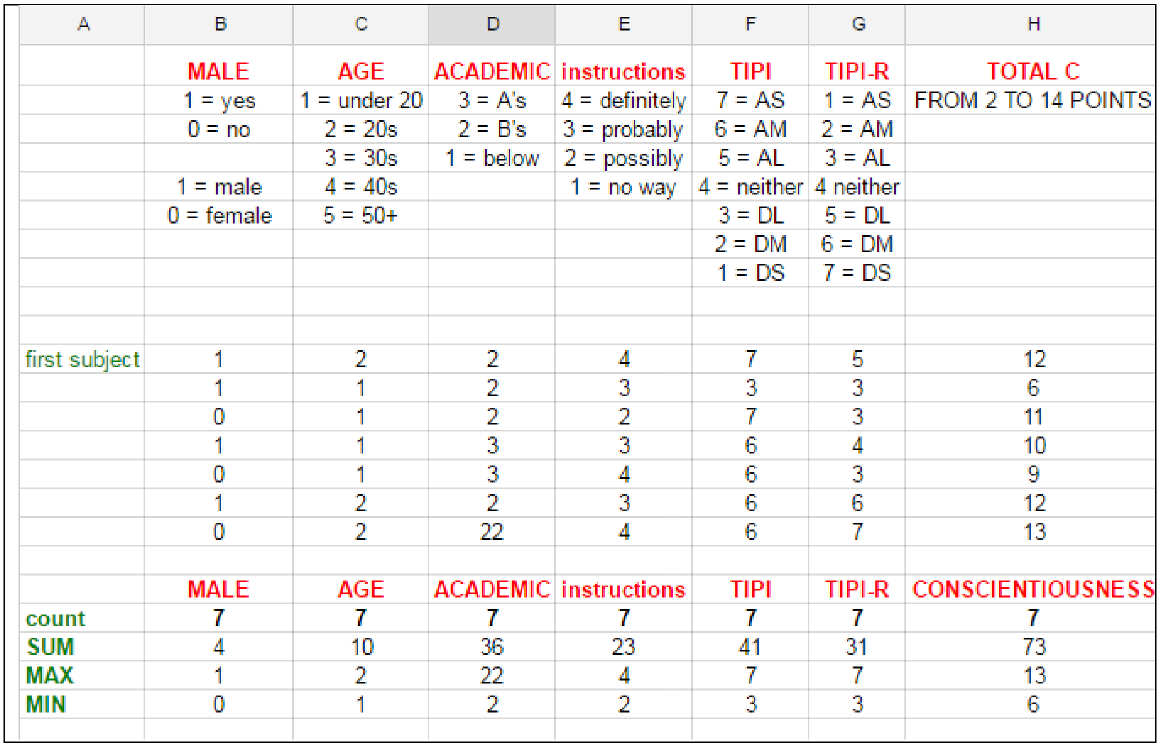
For numerical data, a common measure of dispersion is the *standard deviation,* which is the square root of the variance. This video shows you one of the best ways of calculating the standard deviation. You can also get the standard deviation within a spreadsheet program like Excel or Google Sheets, and within the descriptive statistics section of JASP, Statcato, or SPSS.

However, for error checking purposes, it is best to begin by determining the *maximum*, *minimum*, and *range*. After you finish coding the entire sample, you find the maximum and minimum of each column (a feature built right into Excel and Google sheets). Here's a video on how to use these tests to do preliminary error checking.

For example, if the maximum and the minimum are the same, that means that all the subjects in the sample received the same score. That means that your variable didn't really vary, but was actually a constant. Now, if it was your intention to control that variable, you succeeded. However, if that happens to your criterion variable, then you will be unable to test any of your hypotheses. This kind of result can occur under extreme floor effect or ceiling effect. If it was one of your predictor variables that turned out to be a constant (e.g., only "A" students took the questionnaire) then you will not be able to test any hypothesis involving the predictor variable of academic performance, but you could look at other hypotheses involving other predictor variables.

Any score you entered in a variable’s column that is smaller than the minimum possible value or larger than the maximum possible value is a data entry error. So, suppose you decided to code a criterion variable on a five-level Likert scale one through five. A maximum of 6 means that at least one score was a data entry error. Find which row had the six, and then go back to the original data sheet to get the correct score.

In the example below, look at the variable of academic performance. Notice that the observed range is larger than what the possible coded range was. The maximum should have been no larger than 3, but we observed a 22. We see that the problem of an excessively large value occurred for only one subject (the seventh). So we go back to our *raw data* (the original questionnaires before they were coded) and we look for what subject 7 really said on that item, and see what the real answer was, a B level. So, it looks like when you were entering the data into Google Sheets you just hit the 2 key twice, entering a 22. We correct that figure and notice that the count and maximum rows will be immediately revised, showing that we have numbers within the acceptable boundaries.



A	B	C	D	E	F	G	H
	MALE	AGE	ACADEMIC	instructions	TIPI	TIPI-R	TOTAL C
	1 = yes	1 = under 20	3 = A's	4 = definitely	7 = AS	1 = AS	FROM 2 TO 14 POINTS
	0 = no	2 = 20s	2 = B's	3 = probably	6 = AM	2 = AM	
		3 = 30s	1 = below	2 = possibly	5 = AL	3 = AL	
	1 = male	4 = 40s		1 = no way	4 = neither	4 neither	
	0 = female	5 = 50+			3 = DL	5 = DL	
					2 = DM	6 = DM	
					1 = DS	7 = DS	
first subject	1	2	2	4	7	5	12
	1	1	2	3	3	3	6
	0	1	2	2	7	3	11
	1	1	3	3	6	4	10
	0	1	3	4	6	3	9
	1	2	2	3	6	6	12
	0	2	22	4	6	7	13
	MALE	AGE	ACADEMIC	instructions	TIPI	TIPI-R	CONSCIENTIOUSNESS
count	7	7	7	7	7	7	7
SUM	4	10	36	23	41	31	73
MAX	1	2	22	4	7	7	13
MIN	0	1	2	2	3	3	6

Another common error is to forget to use 1 and 0 for dummy coding (and giving some of the control group members a 2 instead of a 0), or giving out a 0 for the lowest score on an ordinal variable (if we have agreed to score it 1 through 5).

Correlation

Correlation describes the relationship of two variables. This video reviews the essential features of what a correlation is and the key terminology we use for correlations.

A *direct* relationship between two variables (when one is high so is the other; when one is low so is the other) has a positive correlation coefficient. An *inverse* correlation (when one variable is high, the other is low) has a negative correlation coefficient. In this class, positive and negative should not be used to refer to good or bad.

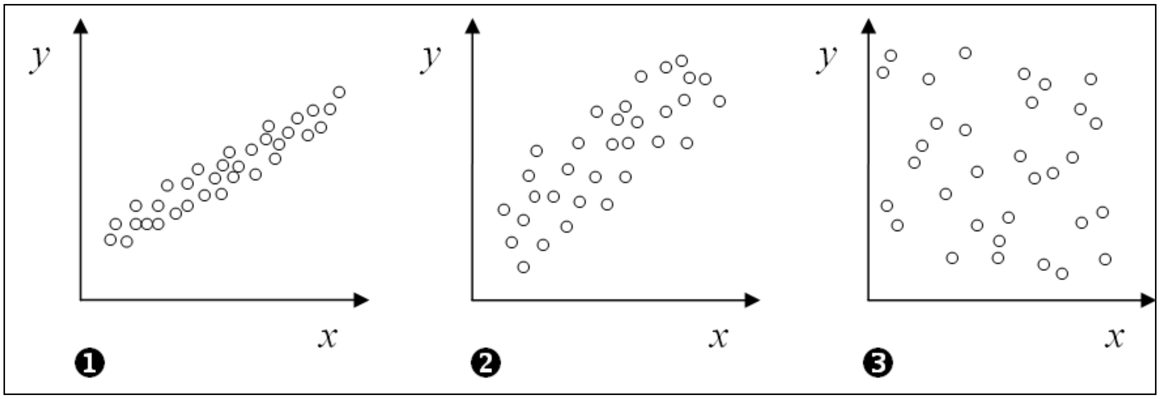
A correlation can be referred to as *strong* (or high) which means that there are very few exceptions to the trend, or *weak* (or low) which means that there are many exceptions to the trend. When there are so many exceptions that no trend can be identified, then there is no correlation (a zero correlation).

A correlation coefficient is a number that has been calculated to show the strength of the relationship between two variables. One commonly used correlation coefficient is the *Pearson* Product Moment Correlation coefficient (symbolized by the small letter r). The Pearson coefficient is parametric, and appropriate when both variables are normally distributed. Nonparametric alternatives include *Spearman's* rho and *Kendall's* tau. These are calculated from ranks rather than raw scores. The Spearman or Kendall are nonparametric coefficients that can be used with skewed interval or ratio scaled data by simply converting the scores to ranks. (Statistical programs like SPSS, Statcato or JASP will do this for you; just tell the program you want to do a Spearman and it will do the calculations right from the numerical data you have entered.) If either of the two variables being correlated is ordinally scaled, or skewed, or truncated, the use of a nonparametric coefficient should be considered because the Pearson coefficient can be greatly distorted in magnitude, and even direction, by one outlier, especially in a small sample.

The closer that coefficient is to zero, the weaker the relationship (i.e., the more exceptions to the trend). Indeed, in a *zero correlation*, there are so many exceptions that there really is no way to identify any trend in the data. A perfect correlation would have a coefficient with an absolute value of 1.00 (which would mean that there is not a single exception to the trend). The closer the coefficient's absolute value is to 1.00, the *stronger* that correlation is.

A bivariate scatterplot is a graphical representation of correlation. Each data point represents one subject. The subjects' place on the graph represents coordinates for two variables. The horizontal position (abscissa) represents variable is X while the vertical position (ordinate) represents variable Y. The closer the data points approximate a straight line, the stronger the correlation.

The scatterplot on the left shows a strong relationship between the variables of X and Y. It would be a positive correlation because those subjects who are high on X also tend to be high on Y. The slope of the regression line is positive: it rises as we go from left to right. The scatterplot in the middle is also for a positive correlation, but one having only moderate strength. Here there are some more exceptions to the trend, though the middle scatterplot also depicts a positive correlation trend. The scatterplot at the right is quite weak, very close to zero, because it is hard to imagine any line approximating the direction of those data points.



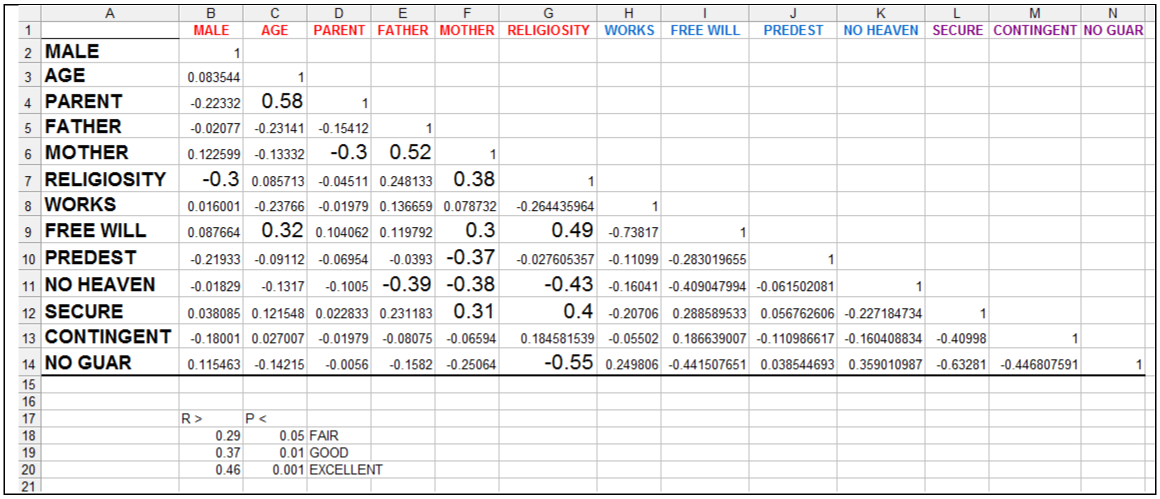
How high is strong? Well, that depends somewhat on the branch of psychology. In neuroscience, we expect higher correlations between variables than we have in clinical psychology, and when we get to social or industrial psychology, the correlations we find sufficient for actionable intervention are even lower. An experimental psychologist focusing on perception might not get too excited about a correlation of -.25, but an industrial psychologist might get excited if she saw that number representing an association between a given predictor variable (a pre-employment test score) and a criterion outcome (a measure of subsequent on-the-job performance).

```
===============================================================
+1.00           perfect positive        no exceptions to trend

        high    strong positive         few exceptions to trend

+.60 ---------------------------------------------------------

                moderate positive       some exceptions to trend

+.20 ---------------------------------------------------------

        low     weak positive           many exceptions to trend

0.00 ---------------------------------------------------------

        low     weak negative           many exceptions to trend

-.20 ---------------------------------------------------------

                moderate negative       some exceptions to trend

-.60 ---------------------------------------------------------

        high    strong negative         few exceptions to trend

-1.00           perfect negative        no exceptions to trend
===============================================================
```

A very useful chart for summarizing the findings of an entire study is a *correlation matrix*. This shows the relationship of every variable to every other variable. Each variable is a row and a column. The number that appears is the coefficient representing the variable of that row with the variable of that column. There will be a diagonal of +1.00 correlations going from the top left to the bottom right (because a variable correlated with itself is +1.00 by definition). Another feature of a complete correlation matrix is that the content of the upper right portion of the matrix (above the diagonal) will be the mirror image of the content of the lower left portion (below the diagonal). This is due to the fact that the correlation between variable X and variable Y is the same whether variable X is the row and variable Y is the column or vice versa. For this reason, some correlation matrices just show the upper right portion of the table (or the lower left portion). This means you might have to look for a variable both in the row and in the column to find its correlation with all other variables.

After a correlation is given, it is important to include some information about its *statistical significance* (e.g., p < .05) or a 95% confidence interval estimate of its range (e.g., rho between +.13 and +.58).

In the example given below of a correlation matrix, the predictor variables included background factors such as gender, age, whether the subject was already a parent, the quality of reported childhood relationships with each parent (mother, father), and current level of religiosity. The criterion variables were the subject's attitudes about salvation (measured on Likert scales). Is one saved by one's own good works, an act of free will acceptance of the savior, predestination, or is there no heaven to hope for? Is one's own salvation secured, contingent upon future behavior, or there is no guarantee of salvation? Correlations between the predictor variables and the first four criterion variables have been enlarged to show them more readily if they were statistically significant.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		MALE	AGE	PARENT	FATHER	MOTHER	RELIGIOSITY	WORKS	FREE WILL	PREDEST	NO HEAVEN	SECURE	CONTINGENT	NO GUAR
2	MALE	1												
3	AGE	0.083544	1											
4	PARENT	-0.22332	0.58	1										
5	FATHER	-0.02077	-0.23141	-0.15412	1									
6	MOTHER	0.122599	-0.13332	-0.3	0.52	1								
7	RELIGIOSITY	-0.3	0.085713	-0.04511	0.248133	0.38	1							
8	WORKS	0.016001	-0.23766	-0.01979	0.136659	0.078732	-0.264435964	1						
9	FREE WILL	0.087664	0.32	0.104062	0.119792	0.3	0.49	-0.73817	1					
10	PREDEST	-0.21933	-0.09112	-0.06954	-0.0393	-0.37	-0.027605357	-0.11099	-0.283019655	1				
11	NO HEAVEN	-0.01829	-0.1317	-0.1005	-0.39	-0.38	-0.43	-0.16041	-0.409047994	-0.061502081	1			
12	SECURE	0.038085	0.121548	0.022833	0.231183	0.31	0.4	-0.20706	0.288589533	0.056762606	-0.227184734	1		
13	CONTINGENT	-0.18001	0.027007	-0.01979	-0.08075	-0.06594	0.184581539	-0.05502	0.186639007	-0.110986617	-0.160408834	-0.40998	1	
14	NO GUAR	0.115463	-0.14215	-0.0056	-0.1582	-0.25064	-0.55	0.249806	-0.441507651	0.038544693	0.359010987	-0.63281	-0.446807591	1
15														
16														
17		R >	P <											
18		0.29	0.05	FAIR										
19		0.37	0.01	GOOD										
20		0.46	0.001	EXCELLENT										
21														

What we see are moderate level correlations indicating the following. Males are somewhat less religious, r = -.30 (because the correlation was negative). Older students (i.e., over age 25) were more likely to be parents themselves (r = +.58) and to agree with the free will acceptance formula for salvation (r = +.32). Those who reported good relationships with their fathers growing up, also reported good relationships with their mothers growing up (r = +.52) and were less likely to deny the possibility of heaven (r = -.39). Having a good childhood relationship with the mother predicted religiosity (r = +.38), the doctrine of salvation by free will (r = +.30) and claiming that one's own salvation was secure (r = +.31). These same subjects who reported good relationships with their mothers growing up were *less* likely to deny the possibility of heaven (r = -.38) or to see predestination as the mechanism for getting to heaven (r = -.37). The greatest predictor of attitudes toward the afterlife was current religiosity. These highly religious people were more likely to claim that their own salvation was secure (r = +.40) and to see free will as the formula for that salvation (r = +.49).

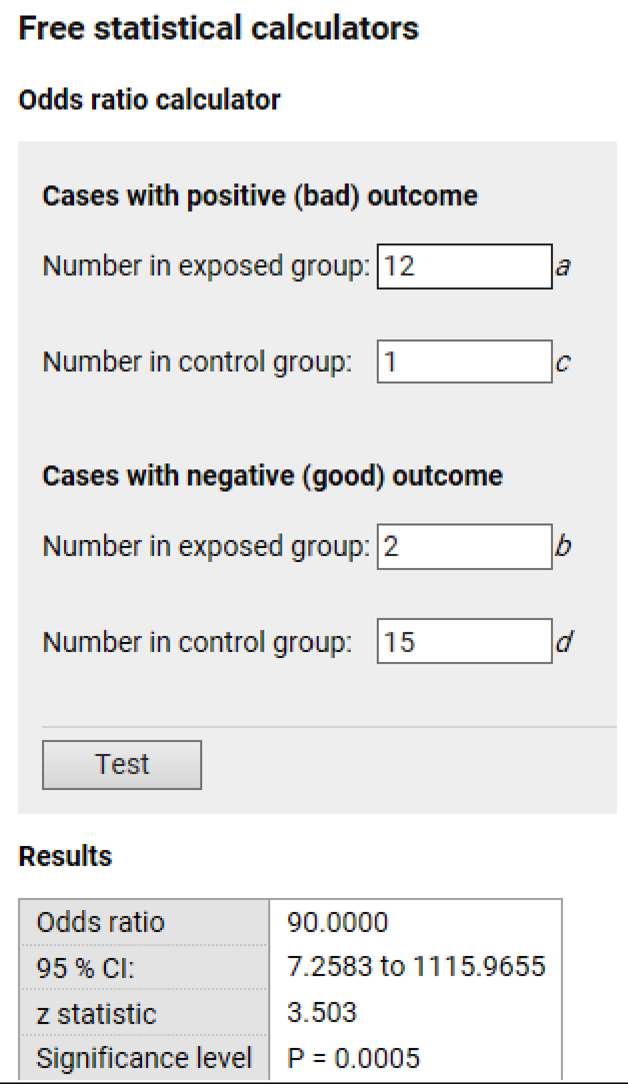
Of course, the above correlations merely show how closely predictor variables are associated with the criterion variables. Causal inference is not certain. We don't know that having a good relationship with one's mother means that she teaches you that you will go to heaven if you accept Jesus. (Indeed, we did not measure exactly what doctrines the mothers taught, and those doctrines probably varied from subject to subject.) Perhaps some other background factor, such as the quality of family life led to many of these outcomes as collateral effects (a spurious relationship): the parental relations, the religiosity and the specific doctrines.

Another way of showing the degree of relationship between variables is *effect size.* This type of calculation is frequently used when the predictor or independent variable is measured on a binary nominal scale (e.g., a variable that sees the subjects in two groups). Effect sizes are frequently used when we do a separate groups comparison, and we get a mean score (and standard deviation) for each group.

There are different effect size coefficients (e.g., Cohen, Glass, Hedges). This site will calculate all three effect sizes. This other site will calculate a Cohen's d from just a mean and standard deviation for each of the two groups (or a t score and the degrees of freedom).

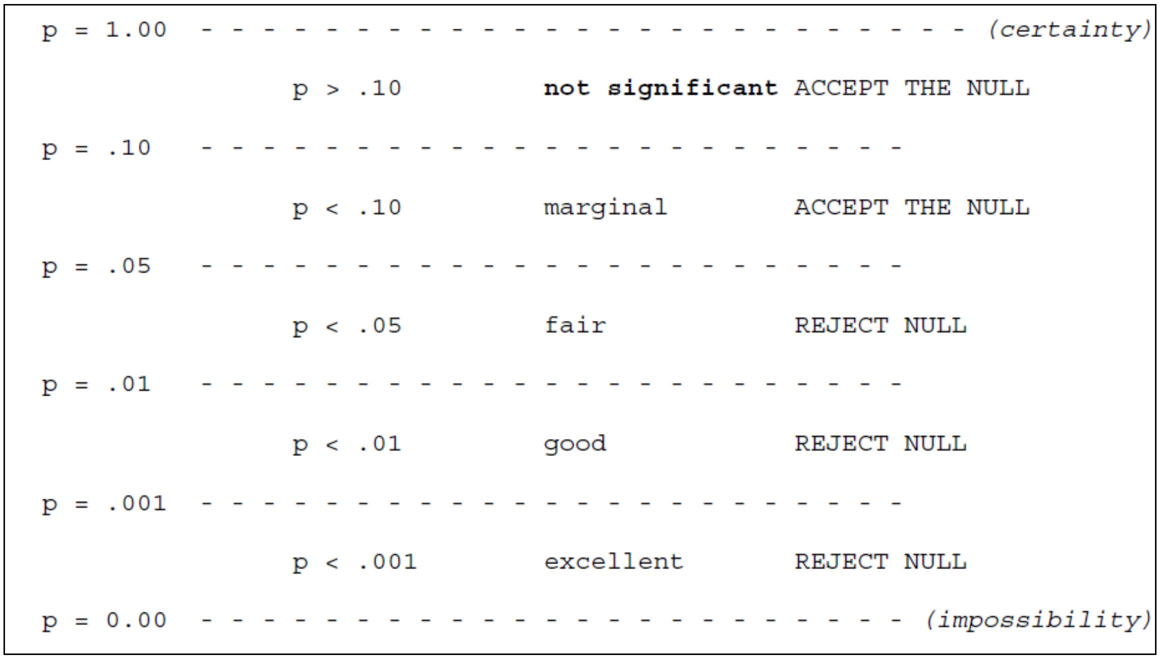
Unfortunately, compared to correlation coefficient strength, there is less agreement about what constitutes a small or great relationship between the variables. Effect sizes do not have a top strength of 1.00, but could theoretically approach infinity. Anything less than .3 would clearly be small, and anything above .8 is generally regarded as large.

Another measure of variable association, especially appropriate for when both variables are measured dichotomously (i.e., in a two-by-two contingency table) is the *odds ratio*. This measure is frequently used in medical research, especially epidemiological studies using a *case control method*. The odds ratio indicates how much more likely one group is than the other to have a certain outcome. So, an odds ratio of 1.00 means no relationship between the variables. The closer the odds ratio is to 0.00 or infinity, the greater the association between the two variables. This MedCalc site will calculate the odds ratio and statistical significance if you can enter the data categorized into the four cells of a two-by-two contingency table.

Notice that in this example, there is a very strong relationship between the variables of exposure (IV) and outcome (DV). Almost all of the exposed cases had a bad outcome, and very few of the control group had a bad outcome, so we have a high ratio (subjects in the exposed group were 90 times more likely to have a bad outcome. Notice that this trend has excellent significance, even for a sample size of only 30. Also note that in medical terminology, *positive* means present (and in this case a disease being present is bad) while *negative* is absent, good.	 **Free statistical calculators** **Odds ratio calculator** **Cases with positive (bad) outcome** Number in exposed group: 12 *a* Number in control group: 1 *c* **Cases with negative (good) outcome** Number in exposed group: 2 *b* Number in control group: 15 *d* Test **Results** Odds ratio 90.0000 95 % CI: 7.2583 to 1115.9655 z statistic 3.503 Significance level P = 0.0005

Inferential Statistics

Inferential statistics are used for determining statistical significance. This video reviews the basic concept. Specifically, inferential statistics calculate or estimate the probability of random variation fitting the results. It is only when we can reject the null (because its probability is less than .05) that we can consider some other explanation for the data. The lower the p value, the *better* our statistical significance. We do not use words like good or bad to describe correlations (we said strong or weak, high or low), but now we are talking about significance and we say excellent, good, fair, marginal or not significant (rather than strong or weak, high or low).

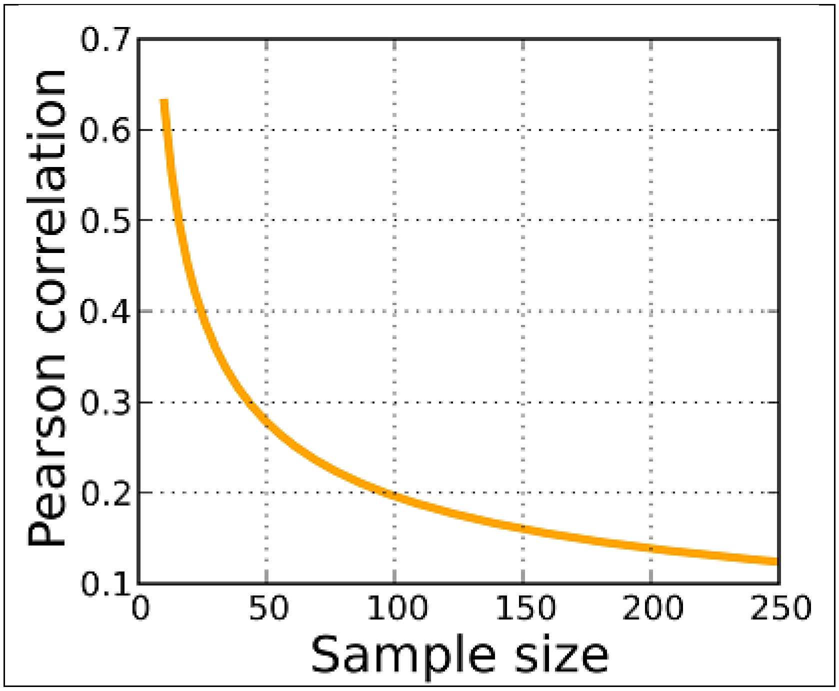


```
p = 1.00  - - - - - - - - - - - - - - - - - - - - - - - - (certainty)

                p > .10        not significant ACCEPT THE NULL

p = .10   - - - - - - - - - - - - - - - - - - - - - - -

                p < .10        marginal        ACCEPT THE NULL

p = .05   - - - - - - - - - - - - - - - - - - - - - - -

                p < .05        fair            REJECT NULL

p = .01   - - - - - - - - - - - - - - - - - - - - - - -

                p < .01        good            REJECT NULL

p = .001  - - - - - - - - - - - - - - - - - - - - - - -

                p < .001       excellent       REJECT NULL

p = 0.00  - - - - - - - - - - - - - - - - - - - - - - - (impossibility)
```

Statistical significance is determined by three factors: the differences between groups, the dispersion within groups, and the sample size. The greater the difference between groups (comparing their means, medians or percents) the better the significance of the difference. The smaller the difference within groups (e.g., a lower standard deviation) the better the significance. The larger the sample size, the better the significance. If you have a small sample size, you will need a large correlation (or a great difference between the groups) in order to have significant data. If you have a large sample size, even a small difference between means may look significant, and that is one reason for the recent trend to include some other measure (such as effect size) in reporting the data.

If the design is repeated measures, replace difference between groups with difference between measurements. If the design is sample vs. norms, the difference refers to the central tendency of the observed in the sample and the central tendency expected from the norms. If the design is correlational, replace difference between groups with strength of the correlation.

Here’s a rule of thumb for the importance of sample size in statistical significance: with a sample size of 50, you need a Pearson r (or Spearman rho) of about .28 (positive or negative) to attain significance at the .05 level. When the sample size is 100, such significance can be attained with a correlation of about .2. The larger the correlation observed, the easier it is to be statistically significant. The smaller the correlation observed, the larger the sample size needed for this difference to become significant.



Parametric statistics include the mean, standard deviation, Pearson coefficient, and inferential statistics such as the t test and ANOVA (and ANCOVA, MANOVA, MANCOVA). These parametric inferential statistics are *powerful* (i.e., they can detect a trend that is even slightly significant) and resist *Type II* error (i.e., accepting the null when we should reject it). These tests can be used when the data are normally distributed or when the sample size is large.

Nonparametric statistics do not assume a normal distribution of the variable, and are appropriate for nominally and ordinally distributed variables, as well as for when numerical data are not normally distributed. Nonparametric statistics include the *test of proportions* (which has both a sample vs. norms version and a separate groups version), *Kolmogorov-Smirnov* (a very robust test, also available in a sample vs. norms version and a separate groups version). There are also exact tests such as the *binomial distribution* (when we are comparing a variable scored on a binary nominal scale to some norms), and the *Fisher Exact Test* (for when we are comparing two groups on a variable measured on a binary nominal scale). The *Poisson* distribution is appropriate for samplings using a given time period, length, area or volume where the dependent variable is in a discrete ratio scale (e.g., number of incidents).

Other examples of nonparametric tests are the *Sign*, *Mann-Whitney*, *Wilcoxon*, *Friedman*, and *Kruskal-Wallis*. Perhaps the best known of these inferential tests would be different formulations of *chi square* (more formally known as chi squared). Here is one of the best sites for its calculation. My verdict is that the chi square is like the “Swiss Army Knife” of inferential statistics. It has dozens of applications (but it is probably not the best tool for any particular task). In other words, there is probably another nonparametric statistic which could be employed whenever you are tempted to use chi squared.

Compared to parametric inferential tests, the nonparametric are less powerful, but more *robust* (resistant to *Type I* error).

In the past few years, there has been growing criticism of traditional null hypothesis testing. Australian statistician Geoff Cummings has led the case for the increased use of *confidence intervals* (indicating a range of possible correlations of effect sizes or correlations within a 95% confidence interval). These seem especially appropriate for large meta-analyses of data. Another alternative, championed by the architects of JASP and APS president C. Randy Gallistel, is the use of *Bayes* factors (which look at the relative likelihood of competing hypotheses, usually a hypothesis more than three times as likely is considered to be significant).

Contingency Tables

The simplest way to analyze categorical data (or ordinal data with few ordered levels) is to cross tabulate with a rows and columns *contingency table*. This is explained in this video.

Use the independent variable (or predictor variable) to define the groups (e.g., rows). Use the dependent variable to indicate what percent of that group has a characteristic (e.g., columns). If you lay it out this way, as you add percents across a row, they should add up to 100% (but this may vary a little due to rounding).

For inferential statistics, you could use a Fisher Exact for a two-by-two table (or a Yates-corrected chi squared). A two-group Kolmogorov-Smirnov could be used for a two-by-whatever table. The rows and columns version of chi squared would work on whatever table, with DF = (rows - 1) X (columns – 1).

A superior way to analyze most data would be to use a spreadsheet program, such as Excel (or Google Sheets), as demonstrated in this video. Use each row in Excel for a different participant (one participant per row, one row per participant). Use each column in Excel for a different variable. The number entered into each cell is the score that subject obtained on that variable. Use a dummy coding (1 and 0). Run a Pearson correlation and it gets pretty much the same result as a chi squared would.

Statistical Spreadsheet Programs

The standard Excel program (and Google Sheets) can do most of the descriptive statistics, a t test, and even a Pearson correlation coefficient between two variables.
To get ANOVA and a correlation matrix, special add-ons may be required for Excel.

For a couple of decades at least, the gold standard in spreadsheet programs remained IBM's SPSS (but the license fees are quite expensive). Other widely used commercial spreadsheet programs include Minitab and SAS.

Now there are some great free alternatives to SPSS. PSPP (an obvious reverse of SPSS) is a GNU program. For usability, I prefer Statcato. For quickness of response and variety of features (e.g., Bayes) the winner is the University of Amsterdam’s JASP. This video shows how to install it.

Regardless of which of these statistical programs is used, my recommendation is that initial data coding be done on Google Sheets (or Excel) and that some descriptive statistics (e.g., count, sum, mean, median, maximum, minimum) be performed as error checks to make sure the data have been entered correctly. Then the data can be pasted into the spreadsheet of SPSS or Statcato. (JASP requires that we open a .csv file in which the first row is the names of the variables and the actual data begin on the second row.)

For the widest range of features, and an open source format that permits continuous innovation, serious programmers prefer the package known as r project. However, most users find r the most complex and least user friendly. So, here's what I use: Google Sheets to code and check my data, followed by Statcato (unless I want to run a Bayesian analysis, in which case I switch to JASP).

In using any calculator, statistical site, or spreadsheet program, realize that the results of the calculations are sometimes reported in *scientific notation* (a.k.a., *E-notation*), especially if the numbers are small (as frequently happens when dealing with significant p values). In physics, astronomy, geology, chemistry and biology, scientific notation is frequently used because the numbers are so large. In psychology, E-values appear as numbers that are very small, so they are negative exponents only. Here are some examples of converting scientific notation back into regular decimals. Remember, when there is a negative sign in front of the E, just move the decimal that many places to the left.

Scientific notation	Decimal notation	Significance
6.4E-02	0.064	Marginal
4.78E-02	0.0478	Fair
1.03E-02	0.0103	Fair
3.21E-03	0.00321	Good
7.32E-04	0.000732	Excellent

Here’s the rule of thumb. E must be at least E-04 to be excellent. This video shows how to use scientific notation.

If the statistical program or does not use E notation for the p values, it is important to be clear whether the decimal number you get is a p value or a test statistic (which you then have to convert to a p value). Excel does not tell you this, but if you ask for a t-test to be performed, the decimal number that you get is the p value, not the t-score.

Tables & Graphs & Charts

One of the nice finishing touches to reporting descriptive and inferential statistics would be the use of tables, charts and graphs.

The contingency table (with the numbers in the cells representing the percent of each group in the cell) may be a good tabular display.

Pie charts can depict the distribution of categories (or levels) of variables. The size of each slice is determined by the percent each slice has.

The *bar* graph can depict raw numbers or measures of central tendency or dispersion. Bar graphs can be used with any scaling of variables.

Line graphs are sometimes used instead of bar graphs, but line graphs are more appropriate for a time series display in which the horizontal variable represents different points in time.

Bivariate scatter plots can depict a correlation between two numerically scaled variables.

There is a chart function within Excel (and SPSS and Statcato and JASP) and this video shows how to use it. However, Excel’s charts are not the most user friendly. JASP charts are easy to get and download, but not always that exciting. An alternative is to do the charts at a simple site, such as this one developed for elementary schools. Then paste these charts into your presentation as a .jpg or .png file. Specific use of these specific graphs will be demonstrated in subsequent chapters.

Which Statistic to Use?

Perhaps the most confusing question is which tool (statistical test) to use. The long answer is it depends on several factors, such as your design (depicted by the rows on the chart below) and the scale on which the criterion variable is measured (depicted by the columns on this graph). For example, if you have a sample vs. norms design, and the variable’s format is multiple nominal, the Kolmogorov is the way I prefer. If you are correlating two ratio scaled variables, but one is skewed (and sample size is small) I would go with the Spearman rho.

Which test to use?

	Binary nominal	*Multiple nominal*	*Ordinal or skewed*	*Interval or ratio*
Descriptive	Percent	Percent, mode	Percent, rank, median, range	Mean, range, standard deviation
Sample vs. norms	Binomial	Chi Squared, Kolmogorov	Kolmogorov	t test
Repeated measures	McNemar Chi	Chi Squared	Sign, Wilcoxon, Friedman	t test ANOVA
Two groups or binary nominal	Fisher Exact, Yates Chi	Chi Squared	Mann-Whitney	t test
Three or more groups or multiple nominal	Chi Squared	Chi Squared	Kruskal-Wallis	ANOVA
Ordinal or skewed	Mann-Whitney	Kruskal-Wallis	Spearman rho	Spearman rho
Interval or ratio	t test	ANOVA	Spearman rho	Pearson r

In practice, regardless of the scoring of the criterion variable, if I have a separate groups or correlational design, I dummy code variables and I put everything into a correlation matrix (of Spearman rho). If some relationship is not significant with that, it probably won’t be significant with some other nonparametric test (e.g., Mann-Whitney) more appropriate for the particular design. If I do detect significance, I might go back and use a more appropriate tool, or even a Bayesian approach, but the correlation matrix is a great first look to see if there are patterns in the data.

Chapter #7: Correlational Designs

Widespread & Easy

Correlational designs are the most commonly used in the social sciences. Indeed, the term is often broadly applied to any non-experimental study in which we are looking at the relationship between different variables, whether the data were gathered by a field count, archival study, questionnaire or trace analysis.

A correlational design does not require any of the following

- Identifiable separate groups for comparison
- Control of any independent variable
- Randomization of any independent variable
- Experimental manipulation of an independent variable
- Repeated measures of any dependent variable
- External norms (e.g., population data) for comparison

In this sense, correlational designs are the easiest to construct and use. We just need two measured variables, and all we have to do is to see if there is some relationship between them (i.e., a correlation coefficient) that can be calculated. Don't worry if you cannot figure out which variable is independent. You can have a correlation between two dependent variables.

Here are some examples of hypotheses that could be tested with a correlational design.

- The *higher* the religiosity, the *lower* the approval of pre-marital sex.
- *Younger* children are *more* likely than older children to show separation anxiety when the parent leaves the room.
- The *higher* a voter's income, the *lower* the likelihood of voting for the Democratic Party.
- *Older* children are more likely to play in *larger* numbers, while younger children prefer playing with a fewer number of peers.
- *High* IQ children will earn *higher* scores on tests of academic achievement.
- Adults who *have* had criminal convictions are *more* likely to have had school suspensions growing up.

Notice that the first three hypotheses state an *inverse* correlation (a negative correlation coefficient) while the last three state a *direct* relationship (a positive correlation coefficient). We could reverse the direction of the hypothesized correlation just by reversing the scoring of a variable. For example, on the last example, adults who *have* had criminal convictions are *less* likely to have had good conduct records during their school years.

The two variables do not have to be scaled the same. One could be interval and the other could be binary nominal, or one could be ordinal and the other ratio continuous. We can nudge Excel, SPSS or Statcato to calculate a correlation coefficient for any of these combinations of scales by using the *dummy* coding (numerical coding of ones and zeros) described in chapter #5.

Limitations

What we cannot deal with is when one of the variables turns out to be a *constant*. For example, if the independent variable is gender and there are only boys in your sample, then you cannot correlate any of the dependent variables with gender (because gender did not vary).

This problem also occurs if the dependent variable turns out to be a constant. Suppose that the intended dependent variable was to be academic performance, operationally defined as whether or not a student passed a test. Suppose that the test was too easy; every participant in your sample passed (a condition known as *ceiling effect*). Now you cannot correlate any independent (or other dependent) variable to whether or not the student passed the test (because that criterion variable is now a constant). The same thing happens on the other end of difficulty, if the test was too hard and nobody passed (a condition known as *floor effect*).

Another thing that correlational designs cannot deal with is when one variable is at the *narrative* level (i.e., representing qualitative data). Correlation coefficients only exist for the nominal, ordinal, interval, and ratio levels of measurement. If your data are at the narrative (or purely visual) level, then they cannot be inserted into an equation for calculating a correlation coefficient. One solution might be to have an expert rater review the qualitative data and assign some categories, levels or numbers based upon his or her interpretation of the narrative. Of course, this opens up questions of precision, validity, and (inter-rater) reliability.

Correlation Matrix

Let’s suppose you were able to quantitatively code five variables: gender, age, grade level, IQ, and academic test scores (and none of them turned out to be a constant). Theoretically, you have ten different correlation coefficients you can calculate (combinations of five things taken two at a time). You don’t have to calculate ten different correlations, one at a time, using a calculator, or a website, or even the standard version of Excel. Most advanced spreadsheet programs (e.g., SPSS, JASP, Statcato, and even Excel with a special add on) can do something called a *correlation matrix*, where with one click you can correlate every variable to every other variable.

Here is what a correlation matrix for these five variables might look like.

	Male	Age	Grade	IQ	Scores
Male	1.00	.06	-.01	-.12	.04
Age	.06	1.00	.87	.10	.38
Grade level	-.01	.87	1.00	.13	.21
IQ	-.12	.10	.13	1.00	.51
Academic Scores	.04	.38	.21	.51	1.00

Each number represents a correlation coefficient between the row variable and the column variable. Notice two things about the correlations in the matrix. One is that as we go along the diagonal, we see 1.00 repeated. That is because any variable correlated with itself gives a perfect direct relationship. Another thing to notice is that the other coefficients are arranged in a mirror image on either side of the diagonal. That is because whether we correlate IQ with scores or scores with IQ, we have the same two variables. It does not matter if IQ is the “x” (row variable) and scores is the “y” (column variable) or vice versa. The calculations will turn out the same. So, some researchers might report an abbreviated version of the above table, something like this.

	Male	Age	Grade	IQ
Age	.06			
Grade	-.01	.87		
IQ	-.12	.10	.13	
Scores	.04	.38	.21	.51

Remember that the numbers in the table are correlation coefficients, and tell us only about the strength (high or low) and direction (direct or inverse) of the association between those two variables, and nothing about statistical significance. JASP, SPSS and Statcato will automatically calculate significance (using a version of the t test) but Excel will not.

To convert any Pearson r coefficient to a p value, go to this site. Scroll down to where it says P from r. The r is the correlation coefficient as a decimal number (don’t worry about the – or + sign). DF stands for *degrees of freedom*. The DF for a correlational design is two less than the sample size (n – 2). The good news is that a correlation matrix is usually calculated on the same sample size, so once you know that a coefficient of .28 (whether positive or negative) is significant to the p = .05 level for your sample size of 50 (DF = 48), then you know that every other correlation in the matrix that is higher than .28 will also be significant.

So, when you report your correlation matrix, you might use asterisks as an indicator of significance, such as

	Gender	Age	Grade	IQ
Age	.06			
Grade	-.01	.87***		
IQ	-.12	.10	.13	
Scores	.04	.38**	.21*	.51***

* p < .05   ** p < .01   *** p < .001

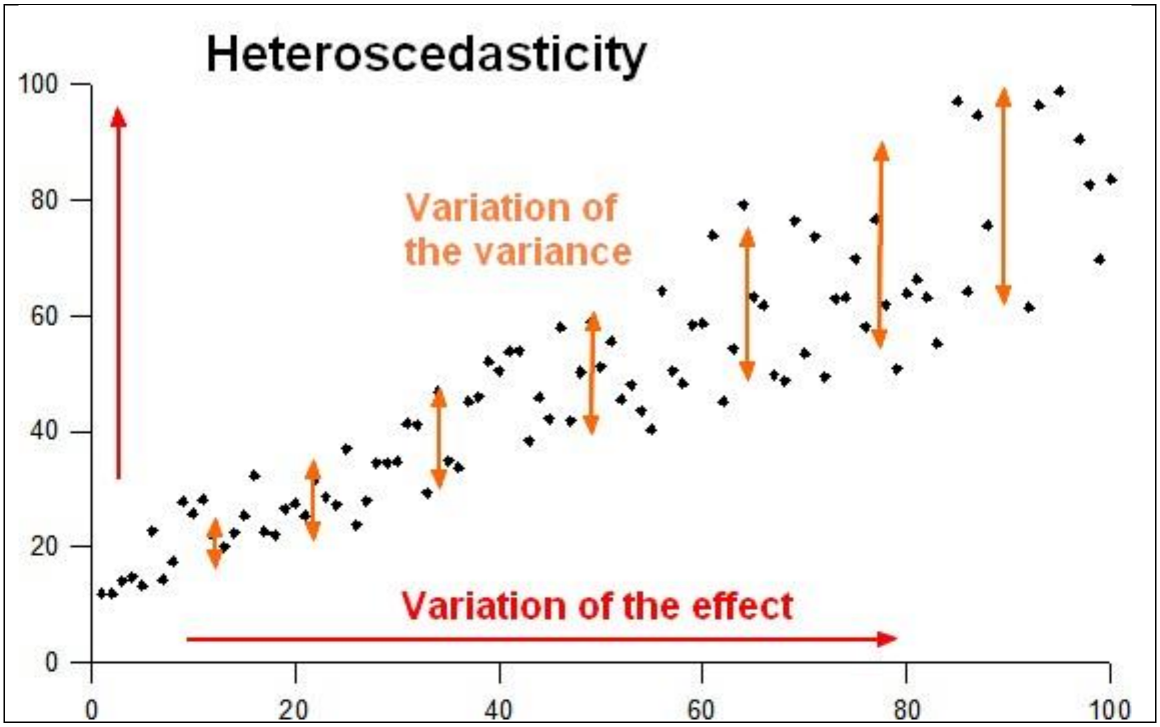
Remember that two factors influence the statistical significance of a correlation: its strength and the sample size. A larger sample size means that we can attain significance with a weaker correlation. A stronger correlation means that we can attain significance with a smaller sample size.

One limitation of most correlation matrices is that these are all just Pearson r coefficients, which only look for *linear* relationships between two normally distributed variables in an interval (or ratio) scale. This term comes from the fact that Pearson r research is usually depicted by a *scatterplot* and a regression line. (Psychologists are usually not very interested in the *slope* and *intercept* of that regression line, but in other areas, e.g., marketing research, it is extremely important because it allows us to *interpolate* (predict certain values of Y for a known value of X within our observed range). The process of predicting a value of Y for a value of X outside of the observed range is known as *extrapolation* and is a much riskier process.

A Pearson r (and *Spearman rho*) will pick up most *monotonic* relationships between variables. These relationships include the linear, and also curved relations as long as it does not change direction. However, the Pearson formula it may exaggerate the strength of nonlinear monotonic relationships, and it could even distort the relationship if there are some outliers and the sample size is small (to the point of saying that a moderate negative correlation is actually a weak positive)! Because of these limitations, it is wise to follow the Pearson calculation with something more appropriate to skewed or ordinal scaling (e.g., a Spearman *rho* or a *Kendall tau* coefficient) or a nonlinear equation (e.g., one that looks at an exponential or logarithmic relationship).

Neither Spearman nor Pearson will detect a *curvilinear* relationship. Such a correlation occurs when the relationship between X and Y is direct over one part of the range of X and then switches to inverse over the other part of the range of X. For example, let X be age and Y be accidents while driving. For drivers under age 50, the correlation between age and accidents is negative: younger drivers have more accidents. For drivers over age 50, the correlation is direct: older drivers have more frequent accidents. If we were to develop a scatterplot relating these two variables, the curve would start high, come down around midlife and then go up again, a U shaped relationship. The best way to deal with such a relationship if it is detected or suspected, is to look for a point along the X axis where that variable can be used to split the sample into two sub-samples (e.g., young and old) and a separate correlation can be calculated between the two variables for each group. If the correlation is curvilinear, the two groups will have correlations of opposite signs.

This same approach of splitting the sample into relevant groups does not always show a curvilinear relationship. Many times what it will show is *heteroscedasticity.* This refers to the fact that the strength of the correlation between X and Y may not be the same over the range of X. Perhaps the direction is the same (direct or inverse), but the strength of that association might vary. For example, there is probably a causal relationship between income and life satisfaction. Over the lower income range, each increase in income produces a corresponding increase in life satisfaction. However, beyond a certain point (it might be $150k, it might be $500k) more dollars of income annually is not going to impact life satisfaction. High income earners are already eating and dressing as well as they can, and more money is not going to allow them to eat more ice cream or wear more clothes. The first vacation home or luxury car does boost life satisfaction, but maybe not the seventh.



Such increased heteroscedasticity may occur even when the correlation is *spurious*: neither X nor Y is causing the other (and both are mere *collateral effects*). For example, Google searches for “autism” and “Asperger” were directly correlated from 2004 to 2012. After the publication of DSM-5 (which redefined autism and eliminated the Asperger diagnosis), there were still Google searches for Asperger, but they no longer had a strong correlation with the amount of Google searches for autism. Both of these searches were dependent variables, but the degree to which they were correlated collateral effects was changed by the publication of DSM-5 (or some other unmeasured factor which occurred in 2013).

Correlation is not Causation

The biggest limitation of a correlational design comes in the area of causal inferences. In general, you can only infer what the causal relationship is between correlated variables if you can do all of the following

- You can reject the null hypothesis (i.e., the data are significant)

- The observed correlation matches the expected direction of the causal hypothesis, e.g., if the independent variable really helped performance, then the correlation must be direct; if the independent variable really hurt performance then the correlation must be inverse. Of course, if the dependent variable is a measure of depression or some other pathology, then if the treatment helped, the correlation would be negative, and if the treatment made the depression worse, the correlation would be positive.

- One of the variables in the correlation is clearly the dependent variable and the other variable is clearly the only independent variable that could have influenced the dependent variable

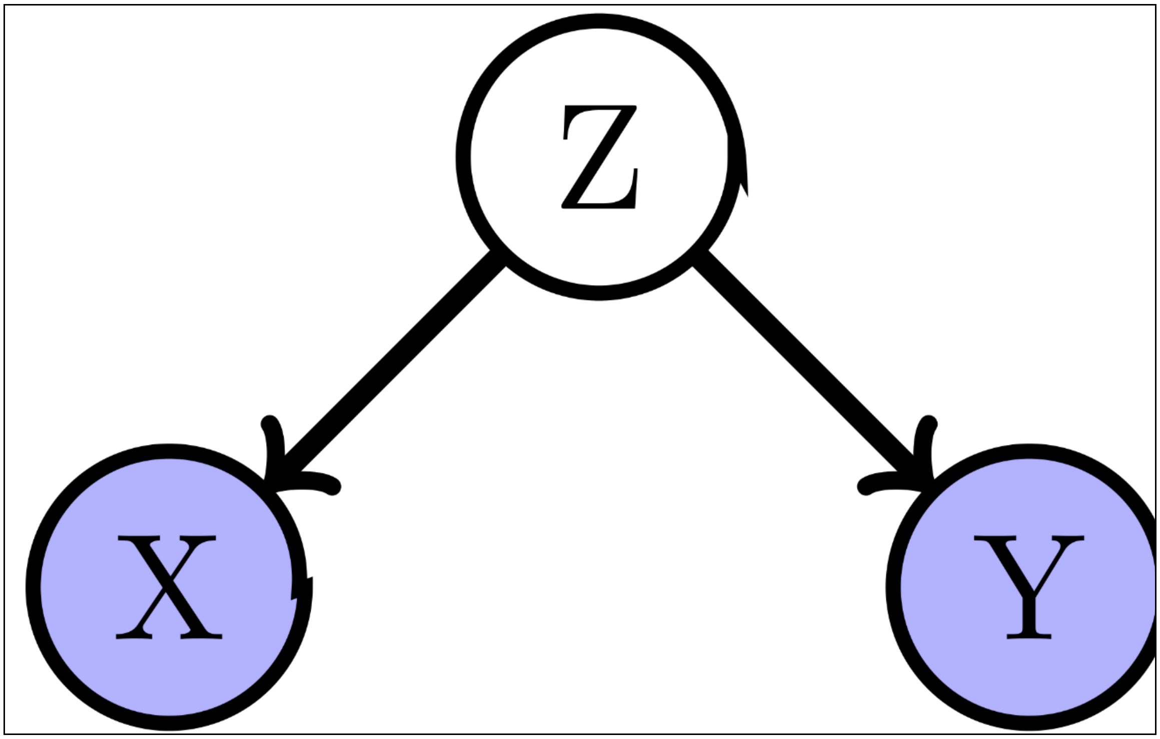
This last criterion is handled best by a true *experiment*, for we manipulate the independent variable and hold constant (or randomize) all other independent variables through random assignment. But correlational designs lack these safeguards, and therefore we often have to say: correlation does not necessarily indicate causation.

We have a special name for this phenomenon, where there are two correlated variables, but neither is the cause of the other. This is known as a spurious correlation. Such a correlation may be based upon precise, reliable and valid measurements of each variable. Such as correlation may be very strong. *Spurious* just means that neither variable caused the other, but both are merely the collateral effects of some other cause (which we may not have measured).

You can still use spurious correlations in this sense: a knowledge of the value of either variable can be used to predict the value of other the other variable for that subject; you just cannot say that one variable caused the other. To falsely assume that we can identify a cause, is a logical fallacy known as *post hoc ergo propter hoc*, as seen in this video about the number of birds outside the Long Beach Performing Arts Center.

The underlying cause of both variables is probably some confounding variable that we did not bother to manipulate, randomize, control, or measure, so we are unable to assess its impact. More about confounding variables can be found in this video.

In this diagram, the correlation between collateral effects X and Y is due to the impact of the confounding variable Z.



If we measured that third variable Z, and correlated it to the other two, we can look at the relative strengths of the correlations, in what is known as *path analysis*. Look at the three pathways possible between the three variables: X & Y, X & Z, Y & Z. Each one of those can have a correlation depicting the strength of the relationship. Path analysis assumes that the weakest of the three paths is not a causal path, and even if it is significant, only represents a spurious relationship between collateral effects.

For example, consider the previous correlation matrix involving a child's age, grade in school, and scores on a standardized test. Yes, there is a positive correlation between the child's grade in school and score on the test, but both of these variables have stronger correlations with the variable of age, and that could lead us to infer that maybe the correlation between grade level and scores is merely spurious.

Chapter #8: Questionnaires

Surveys

The term *survey* is often used synonymously with questionnaire. This book uses the term, survey, more broadly, to cover all scientific research that does not involve actual experimentation. Surveys, in this broader sense, include any non-experimental data collected in the field or a laboratory or stored in archives.

Surveys are not limited to correlational designs. Most surveys do use correlational designs to test hypotheses by seeing if a predictor variable has a relationship with the criterion variable. However, some surveys use separate groups designs. The difference between these surveys and a true separate groups experiment is that the latter manipulates an independent variable, and the survey does not. If the survey is using separate groups, the groupings are due to some background variable (e.g., male/female, older/younger, ethnicity) or current status (e.g., marital, parental, political party, income, residence, employment, denomination) even if that status is the result of the subject's (past or present) decision. If the grouping were randomly assigned, and that led to the two groups being treated differently, then that would qualify as an experiment.

If there are relevant norms for the criterion variable, surveys can use sample vs. norms designs. If we can figure out how to code each subject's data, and compare it to a matched pair (or from the same subject in the future) we can also use repeated measures in surveys.

Not all observations qualify as scientific surveys, just the ones that are systematic and objective. If we only include confirmatory cases in our sample, then we have fallen to the fallacy of *confirmation bias*, and the survey is not scientific. We have to look at cases that are inconsistent with the hypothesis, as well as those that fit, in order to test the hypothesis. This video explores this point in depth.

Types of Questions

What separates questionnaires from other surveys is how the data are collected. Unlike the use of field counts, archives, or traces, when we use questionnaires we require the direct and active participation of the subjects. Questionnaire data are based upon self-reports, and so anything that influences the subjects' ability to comprehend the questions (or respond honestly) may reduce *validity*. Anything that inhibits the subject from responding to a question (e.g., length of the questionnaire) may lead to the great problem of incomplete data.

This book confines the use of the term *questionnaire* to those self-reports involving quantifiable data (e.g., scaling on the nominal, ordinal, interval, or ratio scales). When the questions are *open-ended*, this may lead to a *narrative* level response. This could be considered qualitative research involving either an interview (if there is interaction with the subject) or some kind of textual analysis. Qualitative approaches such as these will be considered in a later unit. In this unit we will confine our discussion to the quantifiable response formats found on questionnaires.

There are over a hundred defects that can plague your questionnaire. This book goes into depth on those problems and provides solutions.

Here are some of the major flaws (which can be fatal).

Using open-ended responses. This is better for an interview situation where you can use questions to initiate a dialogue that will pursue a more in-depth, rich response. If you cannot arrange for a synchronous conversation with the subject, do not use open-ended questions. If you just leave a line or a blank space for the subject to write in, you have no idea if the answer can be quantified in a way relevant to the statistical test of your hypotheses. Even if you had just hoped for a one-word answer like 42 for age, or "carpenter" for occupation or "Baptist" for religion, you might see answers like "old enough" or "construction" or "Protestant" for some subjects, while another subject with the same situation on each of these variables might answer "middle aged, trades, Christian." In other words, if you want a quantifiable response that you can code, you should provide such responses for the subject to select (and have the subject circle the one best response).

Allowing checkmarks. It is sometimes hard to see exactly which alternative the mark is supposed to hit. A better approach is to tell subjects to circle the best answer from several responses.

Non-exclusive alternatives. Maybe the subject finds more than one of the responses acceptable. This happens often with multiple nominal scaling when the responses are not mutually exclusive. This might work with denominational or political affiliation (at least in the U.S.), because it is hard to be a Democrat and a Green at the same time, or a Baptist and a Catholic at the same time. But suppose the question is "What kind of car do you own"? At one time I simultaneously owned a Chevy, a Chrysler, a couple of Lincolns, a Mazda, a Nissan, and several Fords (Mustang, Crown Vic, F-150 trucks). Recently, I have seen this phenomenon apply to ethnicity: students want to circle more than one kind of ancestry. The solution is to change one variable in a multiple nominal scale to several variables with binary nominal scaling even before coding: give the subjects a list of possible answers and encourage subjects to circle each of the categories that applies.

Inadequate alternatives. Another problem is when the subject cannot find the right answer on your list. He owns a different make of car, or belongs to a small denomination, or has not registered with any political affiliation. You might have to include other and/or none on your list. Everyone does not have a career, denominational affiliation, political preference, or a car. Notice, again, how using several variables with binary nominal scaling solves this problem. Someone who does not have a political affiliation will answer "no" when asked if she is a Democrat? Republican? Green? Libertarian? If you do include "other" on your list, realize that such a category includes many (possibly extremely different alternatives). Those who own an "other" make of car include Coopers and Smart Cars, but also Ferraris and Teslas. They don't have much in common except that they don't own a major brand of automobile. Those who belong to the "other" religion would include Wiccans, Zoroastrians and Scientologists. They don't have much in common except that they are not Catholics, Baptists, Jews or Mormons.

Questions that are loaded or leading. These questions build in an argument for one of the alternatives. You will see a lot of these on the "surveys" distributed by your member of Congress asking if you approve of his or her efforts to "control wasteful spending" or "protect national security" or "fight for the middle class." The politician knows that you approve and just wants you to know of his/her efforts. (Did you notice that most of those questionnaires invite you to enclose a contribution to the re-election campaign?)

Ceiling or floor effect. This is where almost all of your subjects answer at one end of the response format. This can occur even without a loaded question. This makes correlations and inferential statistics much harder. Try for a range of alternatives that will get more of a dispersion of the actual responses from the sample. For example, if one of the questions is how often the subject eats out, the following response pattern might be inadequate: more than four times a week / twice a week / once a week / less than once a week. In some geographical locations, among some demographics, people eat out several times a day, and this scale would have ceiling effect. For other demographics, it might have floor effect because some people eat out just a few times a year (or never).

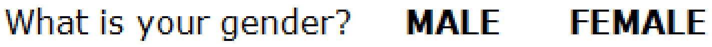
Composite questions. "Do you agree that schools should teach about AIDS and distribute condoms"? There are actually two questions here, and maybe I agree with one, but not with the other. Any sort of global rating (e.g., satisfaction with one's job or marriage) may hide that the subject likes one aspect (e.g., co-workers) but may hate another (e.g., the pay).

The use of branching questions. "If you answered yes, go on to question #4, and if you answered no go on to question #8." You can do this in an interview, or when the questions are being administered orally (in person or over the phone) and the questioner can make the appropriate jump at the right time. You may also be able to achieve this with a computerized administration on a website. However, this is just too confusing for many subjects who are trying to answer a paper questionnaire when the researcher is not immediately present to give guidance.

Specific Questions

This video shows some response formats that generally work.

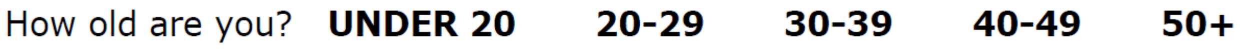
Use this measure for gender. Dummy code as male = 1 and female = 0.



What is your gender? **MALE** **FEMALE**

What is not yet clear is whether the use of a third alternative, such as "non binary" would be called for on this question.

Use this as a measure for age if your population is adult (but not mostly elders). Code under 20 = 1, 20s = 2, 30s = 3, 40s = 4, 50+ = 5.



How old are you? **UNDER 20** **20-29** **30-39** **40-49** **50+**

Perhaps there is some reason why some other intervals might be more appropriate. In a population of high school students, I would put the lowest answer as under the most likely age that someone would attend high school, and the same for the upper bound.

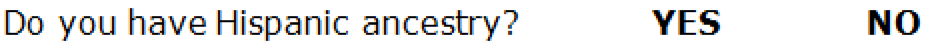
UNDER 14 14 15 16 17 18 19 OVER 20

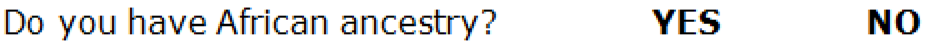
For nursing home residents, I might use this range

UNDER 60 60-69 70-79 80-89 90+

The important thing about selecting the best intervals to use would be that the subject can answer without confusion, and there should be no ceiling or floor tendency. We should also avoid any overlapping categories. So don't say 60-70 and 70-80 because someone who is exactly 70 could circle either category.

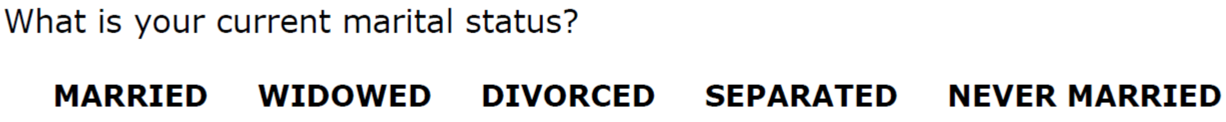
If you want to measure ethnicity, use a series of questions. Do not use one question with numerous categories because many people now identify as having multiple ethnicities. Make each of these categories a separate variable and dummy code each variable as 1 = yes and 0 = no.


Do you have Hispanic ancestry? **YES** **NO**


Do you have African ancestry? **YES** **NO**

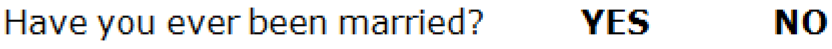
Never ask this question: “Are you single”? That concept is vague and it is unclear how a divorced, separated, engaged, or cohabitating person should answer.

Use the following phrasing if you are interested in current marital status. Make each of these categories a separate variable and dummy code each variable as 1 = yes and 0 = no. So, each subject will get a 1 for one of the variables and a 0 for each of the other four conditions.


What is your current marital status?

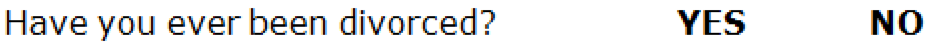
MARRIED **WIDOWED** **DIVORCED** **SEPARATED** **NEVER MARRIED**

Use the following phrasing if what you really want to measure is if the subject has been through the experience of being married. Dummy code as yes = 1 and no = 0.

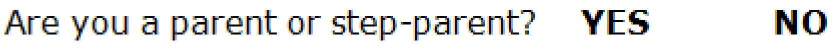

Have you ever been married? **YES** **NO**

On the other hand, if you are dealing with a population in which there is widespread cohabitation without formal marriage you might want to ask about shared living quarters with a fiancé or boyfriend.

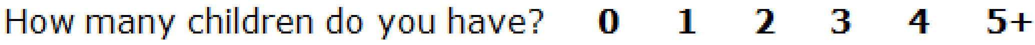
Use the following phrasing if what you really want to measure is if the subject has gone through a divorce. Dummy code as yes = 1 and no = 0.


Have you ever been divorced? **YES** **NO**

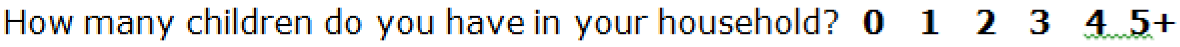
Use the following phrasing if what you really want to measure is whether or not the subject has any children. Dummy code as yes = 1 and no = 0.


Are you a parent or step-parent? **YES** **NO**

Use this phrasing if what you really want to measure the number of children. Code 5+ = 5.

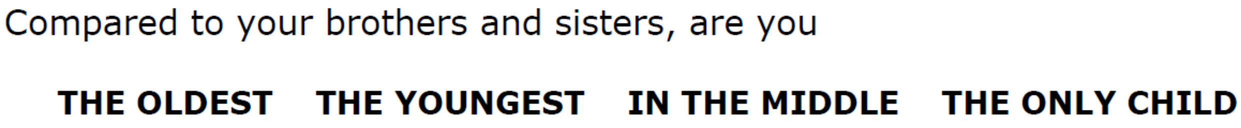

How many children do you have? **0 1 2 3 4 5+**

Use this phrasing if what you really want to measure the number of children that the adult subject is still responsible for. Don’t use this phrasing when some of your population still lives with parents and siblings. Code 5+ = 5.


How many children do you have in your household? **0 1 2 3 4 5+**

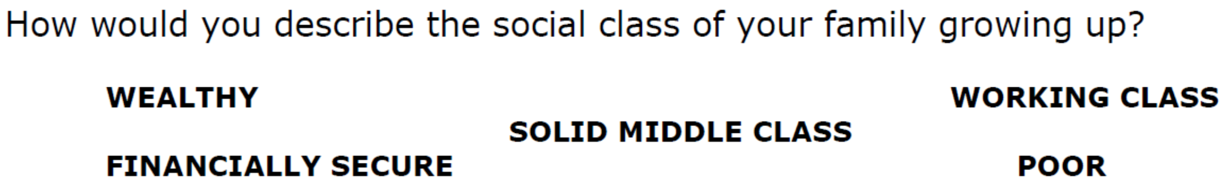
If you think that the term “children” will be confusing (my 89-year-old mother calls me her child) you might want to clarify by using a term like minors or persons under age 18.

Use this phrasing to measure birth order. Make each of these categories a separate variable and dummy code each variable as 1 = yes and 0 = no. So, each subject will get a 1 for one of the variables and a 0 for each of the other three conditions.


Compared to your brothers and sisters, are you

THE OLDEST **THE YOUNGEST** **IN THE MIDDLE** **THE ONLY CHILD**

Use the following phrasing to measure socio-economic class of the family of origin. This is an ordinal scale, and it is important that you correctly code these levels so that the highest score is given to those subjects who score highest on the variable. Code wealthy = 5, financially secure = 4, solid middle = 3, working = 2, poor = 1.



How would you describe the social class of your family growing up?

WEALTHY

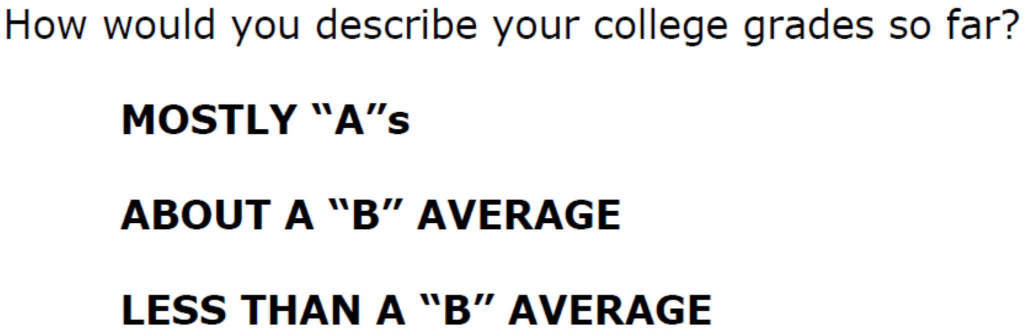
FINANCIALLY SECURE

SOLID MIDDLE CLASS

WORKING CLASS

POOR

Use this phrasing to measure academic performance. Most students do not have a precise recollection of their GPA. It is important that you correctly code this ordinal scale so that the highest score is given to those subjects who score highest on the variable. Code mostly “A”s = 3, about a “B” = 2, less than a “B” = 1.



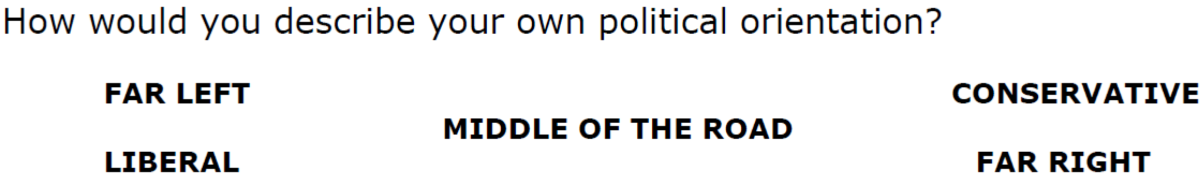
How would you describe your college grades so far?

MOSTLY “A”s

ABOUT A “B” AVERAGE

LESS THAN A “B” AVERAGE

Use this phrasing to measure political orientation. Call the variable “liberal.” This is an ordinal scale, and it is important that you correctly code these levels so that the highest score is given to those subjects who score highest on the variable. Code far left = 5, liberal = 4, middle = 3, conservative = 2, far right = 1.



How would you describe your own political orientation?

FAR LEFT

LIBERAL

MIDDLE OF THE ROAD

CONSERVATIVE

FAR RIGHT

One advantage to using the above phrasing is that we actually have some college student norms for comparison. (You could just complete this table for your sample.)

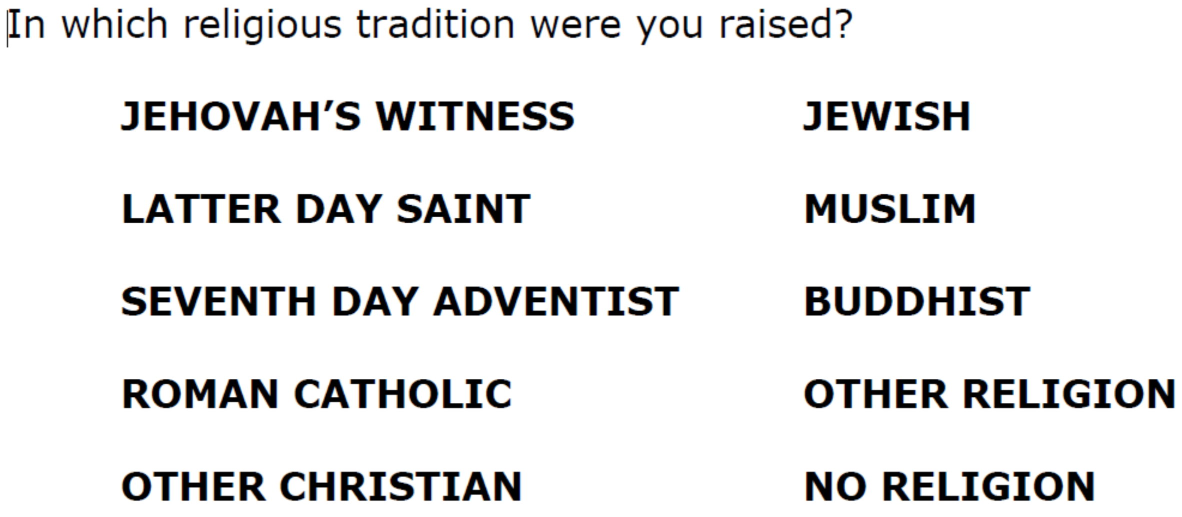
	U.S. Freshmen nationwide	This Sample n	This Sample %
FAR LEFT	3%		
LIBERAL	28%		
MIDDLE OF THE ROAD	46%		
CONSERVATIVE	21%		
FAR RIGHT	2%		

source: Kevin Eagan, Jennifer B. Lozano, Sylvia Hurtado, Matthew H. Case (2013). It can be found here.

One of the most difficult variables to measure on a questionnaire is religion. First, we have to recall the definition of *religion* given in the first chapter. Religion is a system of doctrines, ethics, rituals, myths and symbols for the expression of ultimate relevance. It is best to clarify which component or aspect of religion you want your survey to focus on.

Many times, what you want to investigate is denominational affiliation. Terms to avoid using are “Christian” or “Protestant.” Give a specific list and order it so that the more generic answers are presented later.

We would further have to clarify whether you mean the way someone was brought up or the religion to which the person converted. About a third of Americans change their denominational affiliation at some point in their lives. Denominational affiliation is a multiple nominal scale, so dummy code it as separate variables. Use this phrasing for measuring religion as original denominational affiliation (and make ten columns, coding each subject with a 1 in one of those columns and a 0 in the other nine.



In which religious tradition were you raised?

JEHOVAH'S WITNESS

LATTER DAY SAINT

SEVENTH DAY ADVENTIST

ROMAN CATHOLIC

OTHER CHRISTIAN

JEWISH

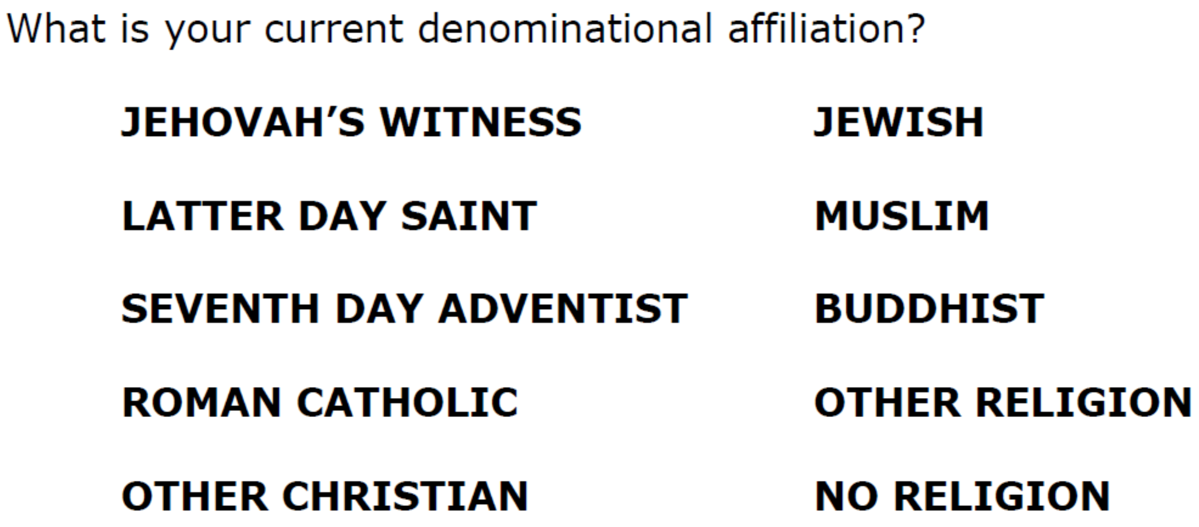
MUSLIM

BUDDHIST

OTHER RELIGION

NO RELIGION

Use this to measure religion as current affiliation.



What is your current denominational affiliation?

JEHOVAH'S WITNESS

LATTER DAY SAINT

SEVENTH DAY ADVENTIST

ROMAN CATHOLIC

OTHER CHRISTIAN

JEWISH

MUSLIM

BUDDHIST

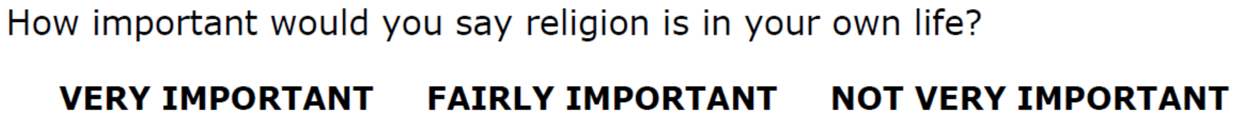
OTHER RELIGION

NO RELIGION

If we had put "Christian" as the first answer the entire first column might have selected that answer. If the fifth answer ("other Christian") or the ninth answer ('other religion") lumps together different traditions that you really want to compare, you may need to come up with more precise phrasing.

If what you really want to do is to compare Baptists and Pentecostals, what you should do is stand outside a local Baptist church one Sunday and outside of a local Pentecostal church on another Sunday. If you just ask people what is their religious tradition, too many will say something like "Christian" or "Protestant" or "Sometimes I attend the Baptist Church, but I have been to the Pentecostals" or "I don't have a religion, it was made by *man*, I just love Jesus because he was sent by *God.*"

Use the phrasing below to measure religion as *religiosity* (the intensity of religious commitment). This is not a measure of how one is religious, but how religious one is. This is an ordinal scale, and it is important that you correctly code these levels so that the highest score is given to those subjects who score highest on the variable. Code very = 3, fairly = 2, not very = 1.



How important would you say religion is in your own life?

VERY IMPORTANT **FAIRLY IMPORTANT** **NOT VERY IMPORTANT**

Here are some norms you may use for comparison.

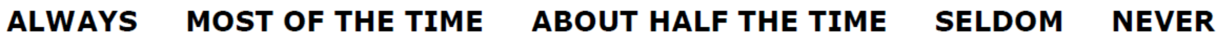
	Gallup Poll (2016)	This Sample n	This Sample %
VERY IMPORTANT	53%		
FAIRLY IMPORTANT	22%		
NOT VERY IMPORTANT	25%		

source: Gallup Poll, 2016, which can be found at this link.

Remember that Gallup's norms are nationwide (including the Bible Belt and Utah) and include all ages. This question can produce ceiling effect if it is distributed at church or Bible study and floor effect if it is used at a secular institution with highly educated Millennials.

One of the most common phrasings of questions involves current frequency of some activity. This could be something that happens to the subject (an independent variable) or something that the subject does (a dependent variable). The following ordinal scaling of responses works for a variety of topics (just replace #### with your specific words. Code always = 5, most = 4, half = 3, seldom = 2, never = 1.

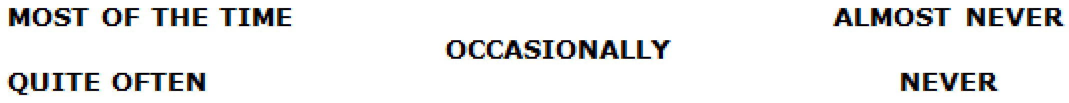
How often do you ################ ?



ALWAYS MOST OF THE TIME ABOUT HALF THE TIME SELDOM NEVER

You can use the following frequency scale to measure things other than depression (just change the words to your particular variable). This is an ordinal scale, and it is important that you correctly code these levels so that the highest score is given to those subjects who score highest on the variable. Code most = 5, often = 4, occasionally = 3, almost never = 2, never = 1.

How often do you feel depressed?



MOST OF THE TIME
QUITE OFTEN
OCCASIONALLY
ALMOST NEVER
NEVER

Here are the norms for the frequency of depression.

Most & Often	10%
Occasionally	45%
Almost never	26%
Never	19%

source: Gallup, G. & Castelli, J. (1989) *The People's Religion: American Faith in the 1990's,* New York, MacMillan, p. 82-83.

Sometimes the variable really deals with events out of the past. Use the phrasing below to measure the self-report of a past frequency: how often something has occurred. Change #### to fit your variable. This is an ordinal scale, and it is important that you correctly code these levels so that the highest score is given to those subjects who score highest on the variable. Code never = 0, once = 1, several times = 2, many times = 3.

How often have you ################ ?

NEVER ONCE SEVERAL TIMES MANY TIMES

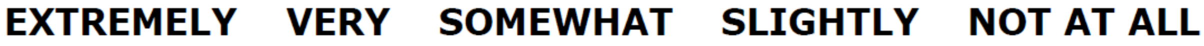
Sometimes we want to measure the participants' perception of how things are going or of people in general. We could also measure the stereotypes associated with a particular group of people. Use this phrasing to measure the subjects' estimates of the proportion of a population having a certain characteristic. Change #### to fit your variable. This is an ordinal scale, and it is important that you correctly code these levels so that the highest score is given to those subjects who score highest on the variable. Code under 10% = 10, 30% = 30, 50% = 50, 70% = 70, over 90% = 90.

What is your best estimate of the percent of ##### who ######## ?

UNDER 10% 30% 50% 70% OVER 90%

Use the phrasing below to measure the **intensity** of interest in something, the relative **importance** of something, or the **appropriateness** of an adjective. Change the #### to fit your variable. This works well as repeated measures with several separate questions, each one asking about the relative importance of something different. Use this approach instead of a forced choice between several options. This is an ordinal scale, and it is important that you correctly code these levels so that the highest score is given to those subjects who score highest on the variable. Code extremely = 5, very = 4, somewhat = 3, slightly = 2, not at all = 1.

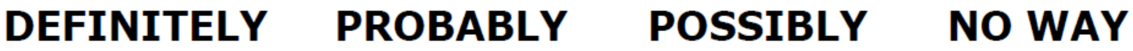
How interested would you be in ##### ?
Would you describe yourself as ##### ?
How important is ######?



EXTREMELY VERY SOMEWHAT SLIGHTLY NOT AT ALL

Use the following phrasing to measure **agreement** with a specific statement. Change #### to fit your variable. This is an ordinal scale, and it is important that you correctly code these levels so that the highest score is given to those subjects who score highest on the variable. Code definitely = 4, probably = 3, possibly = 2, no way = 1.

Would you describe yourself as ################ ?
To what extent is it true that ################ ?



DEFINITELY PROBABLY POSSIBLY NO WAY

Use this seven level *Likert* measure of the subjects' level of agreement with a specific statement. Change the blue

#### to fit your variable. Code as AS = 7, AM = 6, AL = 5, neither = 4, DL = 3, DM = 2, DS = 1.

Do you agree #####?

AGREE STRONGLY		**DISAGREE A LITTLE**
AGREE MODERATELY	**NEITHER AGREE NOR DISAGREE**	**DISAGREE MODERATELY**
AGREE A LITTLE		**DISAGREE STRONGLY**

The above seven-level format is used by the items on the Texas Ten Item Personality Inventory. You could also have a five-level Likert by eliminating the “a little” alternatives.

Use the following item to measure the subjects’ estimate of the **likelihood** of something (in the future). Change the #### to fit your variable. This is an ordinal scale, and it is important that you correctly code these levels so that the highest score is given to those subjects who score highest on the variable. Code VL = 4, SL = 3, SU = 2, VU = 1.

How likely is it that ################ ?

VERY LIKELY	**SOMEWHAT LIKELY**	**SOMEWHAT UNLIKELY**	**VERY UNLIKELY**

You can measure performance, ability, quality or any kind of **evaluation** by the subject using terms like “excellent/good/fair/poor”. Change the #### to fit your variable. This is an ordinal scale, and it is important that you correctly code these levels so that the highest score is given to those subjects who score highest on the variable. Code excellent = 4, good = 3, fair = 2, poor = 1.

How would you rate ####?

EXCELLENT GOOD FAIR POOR

If the question is "How would you rate your level of mental and emotional health?" then here are some norms you may use for comparison.

	Gallup Poll (2001)	This Sample n	This Sample %
EXCELLENT	43%		
GOOD	42%		
ONLY FAIR	12%		
POOR	3%		

source: Gallup Poll Monthly, November, 2001, p. 50

You can measure **performance** by the above phrasing or by using letter grading. Change the blue #### to fit your variable. This is an ordinal scale, and it is important that you correctly code these levels so that the highest score is given to those subjects who score highest on the variable. Code A = 4, B = 3, C = 2, D = 1, F = 0.

How would you rate the performance of #####?

A B C D F

The following measure of comparison to the **average** can be used as a self-rating or to evaluate anything. Change #### to fit your variable. This scale is ordinal, and it is important that you correctly code these levels so that the highest score is given to those subjects who score highest on the variable. Code FA = 5, LA = 4, AA = 3, LB = 2, FB = 1.

Compared to the average #### , how would you rate ####### ?

FAR ABOVE AVERAGE

A LITTLE ABOVE AVERAGE

ABOUT AVERAGE

A LITTLE BELOW AVERAGE

FAR BELOW AVERAGE

Another way to measure performance or satisfaction is with **satisfaction** following terms. Change "your current job" to fit your variable. This is an ordinal scale, and it is important that you correctly code these levels so that the highest score is given to those subjects who score highest on the variable. Code CS = 4, SS = 3, SD = 2, CD = 1.

How satisfied are you with your current job ?

COMPLETELY SATISFIED

SOMEWHAT SATISFIED

SOMEWHAT DISSATISFIED

COMPLETELY DISSATISFIED

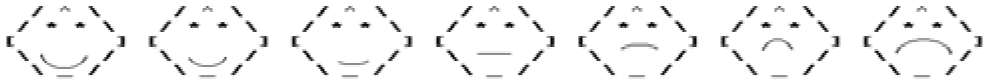
Here are the norms to use with job satisfaction.

	Gallup Poll (2016)	This Sample n	This Sample %
COMPLETELY SATISFIED	54%		
SOMEWHAT SATISFIED	37%		
SOMEWHAT DISSATISFIED	5%		
COMPLETELY DISSATISFIED	4%		

source: Gallup Poll 2016, at this site.

The faces below can be used as a measure of happiness or **satisfaction** with anything. Change the #### to fit your variable. This is an ordinal scale, and it is important that you correctly code these levels so that the highest score is given to those subjects who score highest on the variable. Big smile = 7, medium smile = 6, little smile = 5, no expression = 4, little frown = 3, medium frown = 2, big frown =1.

Mark the face which best demonstrates how you feel about ######### ?



Here are the norms for the face scale when measuring life satisfaction.

	Test Norms	This Sample n	This Sample %
Big Smile	20%		
Medium Smile	46%		
Small Smile	27%		
Flat expression	4%		
Small Frown	1%		
Medium Frown	2%		
Big Frown	0%		

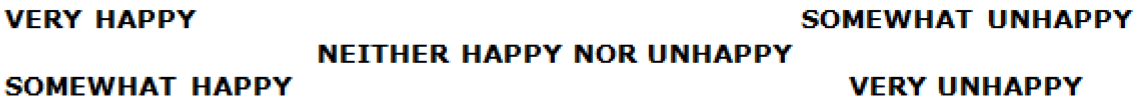
source: Andrews, F.M. & Withy, S.B. (1976)

Social Indicators of Well Being, New York, Plenum.

Kunin, T. (1955) The construction of a new type of attitude measure. *Personnel Psychology, 8*, 65-78.

The following is an ordinal scale, and it is important that you correctly code these levels so that the highest score is given to those subjects who score highest on the variable. Code VH = 5, SH = 4, neither = 3, SU = 2, VU = 1.

How happy are you feeling right now; would you say you are



VERY HAPPY

SOMEWHAT HAPPY

NEITHER HAPPY NOR UNHAPPY

SOMEWHAT UNHAPPY

VERY UNHAPPY

Here are the norms for this measure of happiness.

	Associated Press Poll	This Sample n	This Sample %
VERY HAPPY	21%		
SOMEWHAT HAPPY	42%		
NEITHER HAPPY NOR UNHAPPY	22%		
SOMEWHAT UNHAPPY	4%		
VERY UNHAPPY	9%		

SOURCE: Associated Press, April 16-23 2007, at this site.

Use this numbered scale below as measure of **life satisfaction**. It can also be used as a measure of intensity with 10 = greatest possible intensity and 0 = least possible intensity.

Imagine a ladder. At the top of a ladder is step number 10 representing the best possible life for you and the bottom represents the worst possible life for you. On which step of the ladder do you personally stand at the present time?

0 1 2 3 4 5 6 7 8 9 10

This is known as the Cantril Self-Anchoring Striving Scale.

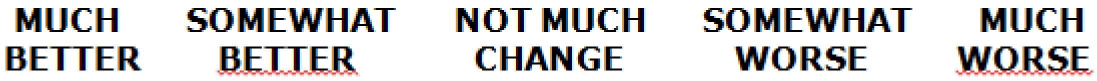
source: Cantril, H. (1965). *The pattern of human concerns*. New Brunswick, NJ: Rutgers University Press.

It has been used widely by Gallup, which has found a median and mode of 7.

Gallup (2009). *World Poll Methodology*. Technical Report. Washington, DC.

Use the following item as a measure of preference for proposed change. Change the #### to fit your variable. This is an ordinal scale, and it is important that you correctly code these levels so that the highest score is given to those subjects who score highest on the variable. Code MB = 5, SB = 4, NMC = 3, SW = 2, MW = 1.

Would it be better if ############### ?



MUCH BETTER	SOMEWHAT BETTER	NOT MUCH CHANGE	SOMEWHAT WORSE	MUCH WORSE

You may have some additional questions for other variables that you need to measure, manipulate or hold constant. Make sure that each response is set off from the question: in a different line, capitalized, bolded, and clearly distinguished from the stem question and the other responses, as in the example below. If you are using an ordinal scale, as in the example below, it is important that you correctly code the levels so that the highest score is given to those subjects who score highest on the variable. In the example below, under a year = 1, between one and five = 2, and over five years = 3.

How long have you been working here at this hospital?

UNDER A YEAR

BETWEEN ONE AND FIVE YEARS

OVER FIVE YEARS

Sampling & Compliance

The biggest challenge with using a questionnaire is getting a sufficient number of subjects and a *representative sample*. The kinds of samples students can usually access do not come close to being *random* (in the sense that every member of the population has an equal probability of getting included in the sample), as explained in this video.

Unless you can justify that your sample is random, do not call it random, but refer to it as a *sample of convenience*.

High levels of *non-response* will usually also *bias* the sample: subjects who care the least about the topics are the ones least likely to return it. To increase response rate, establish a connection with the potential subjects. Also, provide clear instructions on how and when to return the questionnaire.

A related problem of non-compliance is when a subject starts a questionnaire but does not finish. This could be stopping before getting to the end, but a more common problem is skipping one or several of the items.

The best way to deal with this is to eliminate subjects with incomplete data (or at least remove those subjects from calculations involving the missing variable). My own preference would be to remove subjects with any missing data from the entire sample. In other words, when we are coding data on a spread sheet, only complete rows should be included.

Another problem, somewhat harder to detect, is where a subject responds to a long questionnaire by just marking answers rapidly in order to finish (so that the answers are no longer valid answers of a self-report). One way to deal with this is to have some items that are *reverse scored*. For example, the full 30–item version of the Geriatric Depression Scale has twenty items that when answered "yes" indicate depression, but ten items that when answered "no" indicate depression. So, if I see a patient just mark all yes, or all no, I eliminate that questionnaire from the sample, because it is less likely that it was filled out by someone who really scored a twenty (or a ten) than it was by someone who stopped reading the questions.

Other error checkers can look at patterns of response between two questions. For example, if I ask, "Have you ever been married"? And then later "Have you ever been divorced″? The range of possible answers could be yes & yes, yes & no, or no & no, but not no & yes: you cannot go through a divorce if you have never gone through a marriage.

The best way of dealing with the kind of missing and distorted data described above is to prevent it (or at least greatly reduce it) by careful design of these aspects of the questionnaire.

Size matters. Shorter is better. Unless the subjects are highly motivated to finish a long questionnaire, it makes more sense to keep the questionnaire short. My rule of thumb is one side of one piece of paper. That means using short measures of many variables (e.g., one item from Gallup, or two items measuring a trait from the Texas Ten Item Personality Inventory). If you absolutely need to have two pages, make sure and put "turn this sheet over to complete" at the bottom of the first page. (Otherwise about a quarter of participants will forget to do so.)

Visual display matters. If it is very obvious when at item has not yet been answered, it is more likely to get answered. If the response formats differ from item to item, subjects are less likely to fall into a pattern of choosing just the right hand answer for each item.

The Script

This script is the standardized set of instructions that you will use (orally or written on the questionnaire). It is important that your script be standardized in advance for a couple of reasons. First, you want it to be as effective as possible. Second, you need it to be consistent (otherwise this introduces a confounding variable). Researchers who don't standardize their scripts find that the scripts evolve over the course of the study, and vary greatly depending upon the characteristics of the potential participants (e.g., age, gender, appearance).

The first part of the script is the initial approach – your opening words. This is like a sales pitch. You are trying to sell the idea of becoming a participant in the study. When it comes to selling anything, the key should be the needs of the other person, not your own needs. One reason for soliciting participants here on campus is that fellow students, even those who don't know you, are more likely to empathize with you and want to help you complete your project.

Probably the best approach line at Crafton Hills College is "Would you help me with a project I'm doing for Brink's class"? This works better than saying "a psychology class" because some other instructions have a reputation for using really long scales, while Brink has a reputation for short questionnaires about interesting topics. Another effective approach line might be "If you have 90 seconds before your next class starts, you could finish up a questionnaire on the topic of …"

Realize that despite the effectiveness of these opening lines, about half of the students will decline your offer. This is because at the best times (weekday mornings, except Friday) and at the best locations (under the breezeway of the building at the top of the 39 steps) the students are passing through that area at that time because they are going to class (or just getting out of class).

Once a subject has agreed. Give some additional oral instructions, such as "Just circle the best response for each item." Also describe how the subject is to return the questionnaire. Don't allow them to take it home and return it the next day. The questionnaire needs to be returned to you before the student leaves the site (and that is another reason for short questionnaires).

Use a “ballot box” that maintains at least the illusion of security and anonymity. This could be something as simple as a shoe box or cereal box. Show it to each new participant, and encourage him/her to fold up the completed questionnaire and insert it into the ballot box. This is extremely important if the topics are sensitive (e.g., dealing with potentially embarrassing behavior such as number of sexual partners, use of illegal drugs, or diagnosis of mental disorders) the subject might fear that you are looking for personal information about her.

On the questionnaire, at the top, there should be some written instructions on how to fill it out. (This can be a cover sheet if the instructions are long or if there are reasons why we need to get informed consent from the subjects.) At a minimum, you need to say something like

“This is an anonymous questionnaire, so please do not write your name on this sheet. For each of the following questions, please CIRCLE the response that best describes you or your opinions.”

Especially if you are off campus, you need to add a little more to build rapport.

“Hello, my name is ################. I am a #########, but I am also a student at Crafton Hills College. You can help me with a project I am doing for one of my classes.”

So, if you are distributing the questionnaire to law enforcement officers, mention a connection that you might have, such as Police Explorer or member of the Sheriff’s Academy. If you are distributing the questionnaire at the Mormon Church, mention that you are an elder just home from a mission to Brazil. Especially if the questionnaire is to be returned via mail (preferably some kind of internal, organizational mail) you need additional written script at the end clarifying the details of where, when and how. Here is an example.

Please return this completed questionnaire in the envelope provided to the mail box of

***Heather Nguyen** by*

***1 PM Friday, August 18**.*

If you do not have your own mail box there, the best approach is to use one of a respected colleague (rather than someone who might have organizational authority over the subject). In other words, when surveying nurses, use a fellow nurse's mail box, not that of a hospital administrator. When surveying fire fighters, use a colleague's box, not the one belonging to a chief or battalion commander. When surveying a Catholic prayer group, use the mailbox of another layperson, not the priest's.

Important Rules

The most important things to remember from this chapter are the following.

- You are not allowed to distribute a questionnaire until it has been approved by the instructor.

- You are not allowed to change a questionnaire after it has been approved. It is obvious that we should not change the wording of the questions, especially if we are comparing with external norms or a previous administration of the questions. It is obvious that we should not subtract any items (they may be essential in the measurement of a variable required to test a hypothesis). Similarly, do not add any items without permission. It is also important not to change the layout or size of the questionnaire, which happens if you retype it or resize it in some way. (Just print out the camera-ready .pdf file and use it as it is.) If you notice some problem, contact the instructor. You must get his/her approval before you change anything.

- You are only allowed to use subjects from the population that has been approved. In other words, if your approved subjects are “students at Crafton Hills College” you may not increase your sample size by distributing a few questionnaires around the University of Redlands library or to folks waiting at the bus stop. If there arises some great problem in sampling the population originally approved (or some great opportunity to get participants from another population) you must get approval from the instructor before you begin collecting data.

Chapter #9: Randomized Experiments

What is an Experiment?

An *experiment* is research in which an independent variable is *manipulated* (intentionally varied) by the researcher.
In an experiment, the dependent variable must be measured. Do not use any other definition for experiment in this course.

The term “experiment” conjures up images of research done in a laboratory setting. Not all laboratory research is an experiment. If the laboratory is merely used to measure independent and dependent variables, then the research is a form of survey (even if a questionnaire was not used to measure any of the variables). Not all experiments are done in laboratories. Research performed in the field (e.g., a workplace, a marketplace, a hospital) can qualify as an experiment if there is an independent variable manipulated. This video gives non-psychology examples of experiments vs. observational studies.

The requirement for observation to qualify as science is that it be empirical, objective, precise and systematic. Not all sciences use experimentation. Astronomers cannot manipulate variables in dealing with planetary bodies, yet astronomy (not astrology) is a science. Economists can rarely manipulate the variables they study, but economics still strives for the objectivity and precision to qualify as a science. That is the situation with psychology.

Most students equate questionnaires with surveys rather than experiments. However, altering a questionnaire is a very practical way of manipulating an independent variable. Such an alteration could come in the form of different instructions, a different question stem, or a different set of alternative responses from which the subject selects an answer. A common way to do this is to present different hypothetical scenarios (serving as a stimulus, an independent variable) to which the subject must respond (dependent variable).

Here is one example. Suppose we want to find out if the age of a comatose child makes any difference in whether or not people would accept the parents' request to remove the child from a ventilator so that their daughter can "die with dignity." So, we present the scenario:

A six-year-old girl was the victim of a bicycle accident and suffered severe head trauma. The brain damage is permanent and substantial. She is comatose, in a vegetative state, almost certain never to wake up, let alone be able to walk or talk again. She is kept alive only by tube feeding and a respirator that controls her breathing. Although her parents are very religious, they have requested that the child be allowed to pass so that she can go to heaven. Would you grant the parents' request?

DEFINITELY PROBABLY POSSIBLY NO WAY

Imagine that half of the subjects got the above description of the scenario, and the other half had the daughter described as a *sixteen*-year-old. That would be manipulating the variable of the age of the child. The hypothesis might be that people would be more reluctant to let the sixteen-year-old die because society has already invested more in that child.

Or we could manipulate the variable of the child's gender. Imagine that half of the subjects got the above description of the scenario, and the other half had the comatose child described as a six-year-old *boy*. The hypothesis might be that people would be more reluctant to let the little girl die because boys are seen as engaging in more dangerous behavior.

Or we could manipulate the variable of the parents' religiosity. Imagine that half of the subjects got the above description of the scenario, and the other half had the parents of the comatose child described as atheists who regarded the Catholic hospital's attempt to preserve life as an unwarranted intrusion of religious ethics. The hypothesis might be that people would be more reluctant to let the atheists' daughter die because the parents might be perceived as lacking faith.

Or we could manipulate the variable of the parents' socio-economic status. Imagine that half of the subjects got a scenario describing the child as having her bike accident in her gated community (i.e., upper middle class), while the other half of the sample were told that the accident occurred in a poor area of town. The hypothesis might be that people would be more willing to let the poor girl die because she would not be likely to do better than a life of poverty even if she did live.

So, what makes something an experiment is not whether it is done in a laboratory, in the field, or even using a questionnaire. What makes something an experiment is the manipulation of an independent variable.

Experiments are the best research method for inferring a causal relationship between the variables. Mere surveys using a correlational design are prone to *spurious* correlations and the *post hoc fallacy* that "variable X caused variable Y" just because the two variables are associated. The true experiment gets around this post hoc problem by manipulating the supposed independent variable and then seeing if there is any resulting change in the dependent variable.

Dealing with Confounding Variables

In addition to manipulating the independent variable, and measuring the dependent variable, there is another requirement for a well-done experiment. All other potential causes must be accounted for. Anything else that could influence the dependent variable must be measured, controlled, or randomized. These other potential independent variables are known as *lurking*, *extraneous* or *confounding* variables, as seen in this video.

Surveys handle these variables by trying to measure as many as possible, correlating them to the criterion variable, and seeing if the resulting correlations are lower than the correlations between the variables which are the focus of our research. If these other variables have weaker correlations, the assumption is that they are probably not the major causes of the dependent variable.

Most experiments are done with a *separate groups* design, which is also known as a *between-subjects* experiment. This is especially true for an experiment about the impact of treatment (e.g., psychiatric medication, training). These two groups are usually called the *experimental* (which receives the treatment) and the *control* (which does not receive the treatment, and is used as a comparison for the experimental group). Notice that the latter group is known as the *control* (and not control**led**). If the independent variable is something other than a treatment or some other all or nothing phenomena, the two groups might represent different levels of the independent variable. For example, if the independent variable was temperature, one group might be in a warm environment while the other would be in a cooler environment.

While other research using separate groups (e.g., surveys) merely identifies pre-existing separate groups (e.g., male/female, old/young, Republican/Democrat, Mountaineers/Flatlanders) the true experiment uses *random assignment* to the groups, as described in this video. Indeed, another name for this kind of experiment is randomized control trial (RCT).

Remember that the word random means equal probability. When we spoke of random sampling, we meant that each subject in the population had an equal chance (compared to every other subject) of being selected into the sample. Random assignment means that each subject in the sample has an equal chance (compared to every other subject in the sample) of being assigned to the experimental group. For example, if the sample size is n = 50, and we can assign 20 to the experimental group, and the remaining 30 will serve as controls, the assignment would be random if each subject in the sample has that same 40% chance of making it into the experimental group, and no other factor like the subject's gender, age, ethnicity, performance or preference can raise or lower that probability.

This kind of *randomization* is an effective way of dealing with potentially confounding variables (especially when the sample size is large). For example, suppose I am doing an experiment to see if using online drills (independent variable) improves student performance in the course (dependent variable). Some of the students are randomly selected to get access to the drills (the experimental group) while the rest are not (the control group). Think of all the other factors that might influence student performance in the class: IQ, motivation, outside work, supportive parents, study skills, previous coursework. These are potentially confounding variables, and the reason that we would use random assignment.

If we asked the students to volunteer for the grouping, the most motivated students would probably opt for doing the drills (in addition to studying in other ways). So, if we found that the experimental group ended up doing better on the final exam (the operational definition of the dependent variable) would we attribute this to the drills or the initial motivation that got them to try to do better by using the drills?

Random assignment means that the main difference between the groups will be the treatment condition (in this case the drills), and that motivated students are likely to end up in more or less equal proportions in the two groups. The same will happen to the other extraneous variables (e.g., previous coursework, supportive parents). Of course, it is possible that all the better prepared and motivated students will disproportionately end up in one group, but the larger the sample size, the less likely such an unequal distribution would be.

The other way of dealing with extraneous variables is to control them, and that means to turn them into *constants*. If we are concerned that gender might have an impact on the results, maybe we should just include females in our sample for this round of research. Therefore, both the control and the experimental groups will have only females. In this way, male gender could not impact the results. If we are concerned that age might be a factor, maybe we should tighten the age range for our sample to 18-22. If we think that previous coursework is a factor, we could just include data from those students who have already passed Math 108.

Another way of controlling a variable would be to intentionally assign a proportionate amount of persons with that variable to both the experimental group and control group. If the entire sample was 55% female, we could make sure that both the experimental group and the control group ended up with 55% female. If only 30% of students in the entire sample had passed Math 108, we could make sure that both the experimental group and the control group had that same proportion of prepared students. If the average

age of the sample was 22 years, we could assign students to the two groups in such a way that the groups would have the same average age of 22. With these approaches to control, it could not be said that either group would have an advantage on the other in terms of age, gender or academic preparation, and therefore we would be more confident in our inference that it was the independent variable manipulated (the drills) accounting for any observed difference in the performance of the groups.

Almost a hundred years ago, a classic industrial psychology experiment was done at a plant in Hawthorne, Illinois. The researchers manipulated a series of independent variables (e.g., lighting, table height) and observed higher and higher rates of production (the dependent variable). Since there was no real control group, the researchers began to wonder if the changes in worker performance could be attributed to the mere fact that the workers knew they were being monitored. The term *Hawthorne Effect* has come to stand for the explanation that the subjects' very knowledge that they are being observed may change their behavior.

Another related factor boosting the performance of the experimental group is that if they know they are receiving the treatment, this may elevate their expectations and lead to improvement (a sort of self-fulfilling prophecy). On the other hand, if the subjects in the control group know that they are not receiving treatment, this may make them more pessimistic and may impair their performance (or recovery from depression).

In studies of psychiatric medication, a *double blind* format is used to deal with the possible influence of the Hawthorne and expectation effects. We might take a sample of patients composed exclusively of patients who are clinically depressed (so their starting point is controlled in the sense of being held to a tight range). We randomly assign some of the sample into an experimental group receiving a new anti-depressant medication. To manage the control group's expectations, they will receive a *placebo* (a pill that does not contain any active ingredients). So, all of the patients are taking a pill, and all are having their levels of depression monitored. None of the patients know if they are getting the real medication or the placebo. Additionally, the nurses who are handing out the pills and monitoring the patients' levels of depression don't know which patient is getting what (so both patients and researchers are “blind” to what is really going on). This double blind approach is supposed to control both the patient and staff expectations about which patients are supposed to improve.

Another way to control for extraneous variables can be found in sampling procedures. With laboratory experiments, you have to recruit volunteers. The more homogeneity in the sample, the fewer the confounding variables (but the less the sample is representative of the population). With field experiments, the big questions are which potential subjects are selected into the sample (and how can we truly randomize who gets the treatment).

For example, each semester I usually get one student who wants to do a field experiment in which she will smile at half of the people she makes eye contact with, and the dependent variable is whether or not the subject smiles back. We have to make sure that the researcher does not just choose to smile at those who look the friendliest to begin with (or the ones who are better looking, or just the ones that she already knows).

Another related control factor is to standardize the procedures. Each subject should receive the same instructions and the same measurements of the dependent variable. Examiners have to be trained in how to rate the subjects’ performance or improvement. Inter-rater reliability of these assessments needs be established, not just assumed. It really helps to have a run through (a pilot test) with a small number of subjects not counted in the later sample before we finalize standardization and begin data collection.

One particular area deserving attention in the *pilot* test of a separate groups experiment is the adequacy of the manipulation of the independent variable. We have to avoid a situation where the difference between the experimental and control conditions is too little (e.g., the dosage of the medication is too low).

We speak of an experiment as having *internal* validity to the extent that it has accomplished these tasks: validly measured the dependent variable, adequately manipulated the independent variable, and used randomization and control to eliminate possible confounding variables. If so, then the experiment can accomplish its goal of inferring whether changes in the dependent variable can be attributed solely to the manipulation of the independent variable.

We speak of an experiment as having *external* validity to the extent that this causal inference can be generalized outside of this particular study. Sometimes the laboratory procedures which have been used to control all the extraneous variables have created such an artificial environment that it simply does not reflect what is found in the real world. External validity is an important point to bring up toward the end of the discussion section of your write-up.

Coding & Statistics

When it comes to coding the data, the best way is to use a spreadsheet (e.g., Excel, Google Sheets) with each subject occupying a separate row and each column a different variable. The treatment is one variable (column). If you are intending to use a big correlation matrix, then dummy code this variable. Code the experimental group as a 1 and the control group as a 0 (that way, when you see the correlations, a positive correlation will indicate that the experimental group had a higher score on the dependent variable, and the control group had a lower score). Here is an example of how that would look, assuming that the dependent variable is an attitude that we hope to influence by an experimental treatment (such as an advertisement that the experimental group was shown but the control group was not shown). Here, we measure the dependent variable on a five-level Likert scale of level of agreement with a statement.

	GROUP	MALE	AGE	HISP	MARRIED	LIKERT
	1 = exp	1 = yes	1 = over 25	1 = yes	1 = yes	5 = strongly agree
	0 = control	0 = no	0 = under 25	0 = no	0 = no	4 = mostly agree
						3 = not sure
		1 = male			1 = married	2 = mostly disagree
		0 = female			1 = widowed	1 = strongly disagree
					1 = divorced	
					1 = separated	
					0 = never married	
first subject	1	0	0	1	0	3
second subject	1	1	0	1	0	4
third subject	0	1	0	0	0	4
fourth subject	0	1	0	0	1	5
fifth subject	1	1	1	1	0	2
sixth subject	0	0	1	1	0	4

Using just the above six subjects as our example, the descriptive statistics for the sample would be below.

	GROUP	MALE	AGE	HISP	MARRIED	LIKERT
SUM	3	4	2	4	1	22
MEAN	0.5	0.67	0.33	0.67	0.17	3.67
MEDIAN	0.5	1	0	1	0	4
MODE	1,0	0	0	1	0	4
MAX	1	1	1	1	1	5
MIN	0	0	0	0	0	2
STD. DEV	0.55	0.52	0.52	0.52	0.41	1.03

With the above data we see that there were three in the experimental group (so we just subtract that from the sample size of six to find out that there must have also been three in the control group. We see that 50% (a mean of .5) of the sample was in the experimental group, so we subtract 100% - 50% = 50% to find out that we had 50% in the control group as well. (Remember, the experimental and control groups do not have to be equal in size. We are comparing measures of central tendency such as means and percents that adjust for the different group sizes.) We see that there were four males (for 67% of the sample), and only two students (33% of the sample) were over age 25. The majority of this sample (67%) was Hispanic, and only one person (17%) was married. On the Likert scale, our dependent variable, we had a range of scores from two to five (meaning that no one in this sample selected the lowest possible answer of "strongly disagree"). The modal (most frequent) and median (middle) response of the entire sample was "mostly agree" and if we go by the numerical codings of the outcome variable, we could express the average as a mean of 3.67.

If you want the spreadsheet (Excel or Google Sheets) to help you calculate the number in the experimental group (and control group), the percent in each group, and descriptive statistics on the outcome variable (mean, median, mode, maximum, minimum, standard deviation), it would also help to enter all members of one group together before you enter any members of the other group. For example, if you have 50 subjects in the sample, (experimental = 23, control = 27) and start entering data on row 11, then rows 11 through 33 will be occupied by the experimental group and rows 34 through 60 will be occupied by the control group. We can then run such descriptive measures for central tendency and dispersion on each group as well as the entire sample.

If we know that we are not going to use a correlation matrix to analyze all the variables together, we would not have to dummy code the treatment variable. We could use labels such as EX and CN all down the column to indicate assigned group. This would have the advantage of greater clarity in tables generated by the programs.

This might be the point at which we move from a simple spreadsheet program to a more sophisticated statistical package, by copying and pasting our data into the rows and columns of a program like SPSS or Statcato or JASP. With SPSS and Statcato, there is a special row above the spreadsheet for us to paste in the name of the variables and then we can paste in all the numerical data (e.g., rows 11 through 60), but on the statistical program it will occupy rows 1 through 50.

JASP will not allow us to paste in the data. We have to open the data as a .csv file. The top row should be the variable labels and subsequent rows (2 through 51) would be the data for the subjects. When JASP opens the file, data appear in rows 1 through 50. (Remember to do any data editing in Excel or Google sheets before you open it with JASP.)

We could show the relationship of the independent variable and the outcome variable as a (*point biserial*) correlation coefficient. If we put all the dummy coded variables into a

JASP Pearson correlation matrix, with the outcome variable as our first variable, then the top row of that correlation matrix would show the correlation of our outcome variable with every other variable, not just with the manipulated independent variable (i.e., the groups). Assuming that the criterion variable is entered in the last column of the data spreadsheet, the last column of the JASP matrix will show its correlation with each of the other variables as predictors.

If we just have two groups (e.g., experimental and control) and one criterion variable measured in another binary nominal scale (e.g., pass/fail, yes/no) we could use the old, less sophisticated two-by-two contingency table to tabulate and report our data.

Contingency Table for Separate Groups			
	DV = pass	*DV = fail*	*Totals*
IV = yes (experimental group)	A	B	A + B
IV = no (control group)	C	D	C + D
Totals	A + C	B + D	N = A+B+C+D

Cell A represents those subjects in the experimental group who passed the performance test.

Cell B represents those subjects in the experimental group who failed the performance test.

Cell C represents those subjects in the control group who passed the performance test.

Cell D represents those subjects in the control group who failed the performance test.

The marginal A + B represents all those in the experimental group.

The marginal C + D represents all those in the control group.

The marginal A + C represents all those in the entire sample who passed the performance test.

The marginal B + D represents all those in the entire sample who failed the performance test.

A + B + C + D will represent our N, the sample size.

In doing error checking, remember that the sum of the horizontal marginals must equal the sum of the vertical marginals which must equal the sample size.

If the same proportion of the first row is in the first column (compared to the proportion of the second row in the first column) then there was no impact of the independent variable on the outcome.

No impact of the Treatment on the Outcome			
	DV = pass	*DV = fail*	*Totals*
IV = yes (experimental group)	10 67%	5 33%	15 100%
IV = no (control group)	20 67%	10 33%	30 100%
Totals	30	15	**N = 45**

In the above example, two-thirds of the subjects passed the test, whether or not they had received the treatment.

Therefore, the treatment had no impact on the outcome.

The Treatment **Helped** the Outcome			
	DV = pass	*DV = fail*	*Totals*
IV = yes (experimental group)	12 80%	3 20%	15 100%
IV = no (control group)	5 14%	30 86%	35 100%
Totals	17	33	**N = 50**

In the above example, 80% of the experimental group passed, but only 14% of the control group was able to do so without the treatment. Therefore, the treatment helped subjects to pass the test.

The Treatment **Hurt** the Outcome			
	DV = pass	*DV = fail*	*Totals*
IV = yes (experimental group)	10 40%	15 60%	25 100%
IV = no (control group)	25 50%	25 50%	50 100%
Totals	35	40	**N = 75**

In the above example, 40% of the experimental group passed, but half of the control group was able to do so without the treatment. Therefore, the treatment hurt subjects' performance on the outcome measure.

In order to find out if the degree of helping or hurting is statistically significant, we have to factor in the difference between the two groups and the size of each group. The proper inferential statistical test to use with the above contingency tables would be the Fisher Exact Probability or the Yates version of Chi Squared.

Tables & Charts for Separate Groups

The most common situation for a separate groups design experiment is where we have two groups, and we are comparing them on some dependent variable measured on some interval or ratio scaling. Let's take the following example. The population is depressed elders, with a sample of 74. This psychiatric experiment involves a new SSRI anti-depressant medication tested against placebo. The outcome variable is the 30-item Geriatric Depression Scale. Results could be summarized in this kind of table with two measures of central tendency (mean and median) and both range and standard deviation as measures of dispersion.

Group	N	Mean	Median	Maximum	Minimum	S.D.
Ex	24	9.6*	8	18	3	4.6
Cn	50	13.6*	12	27	7	5.2

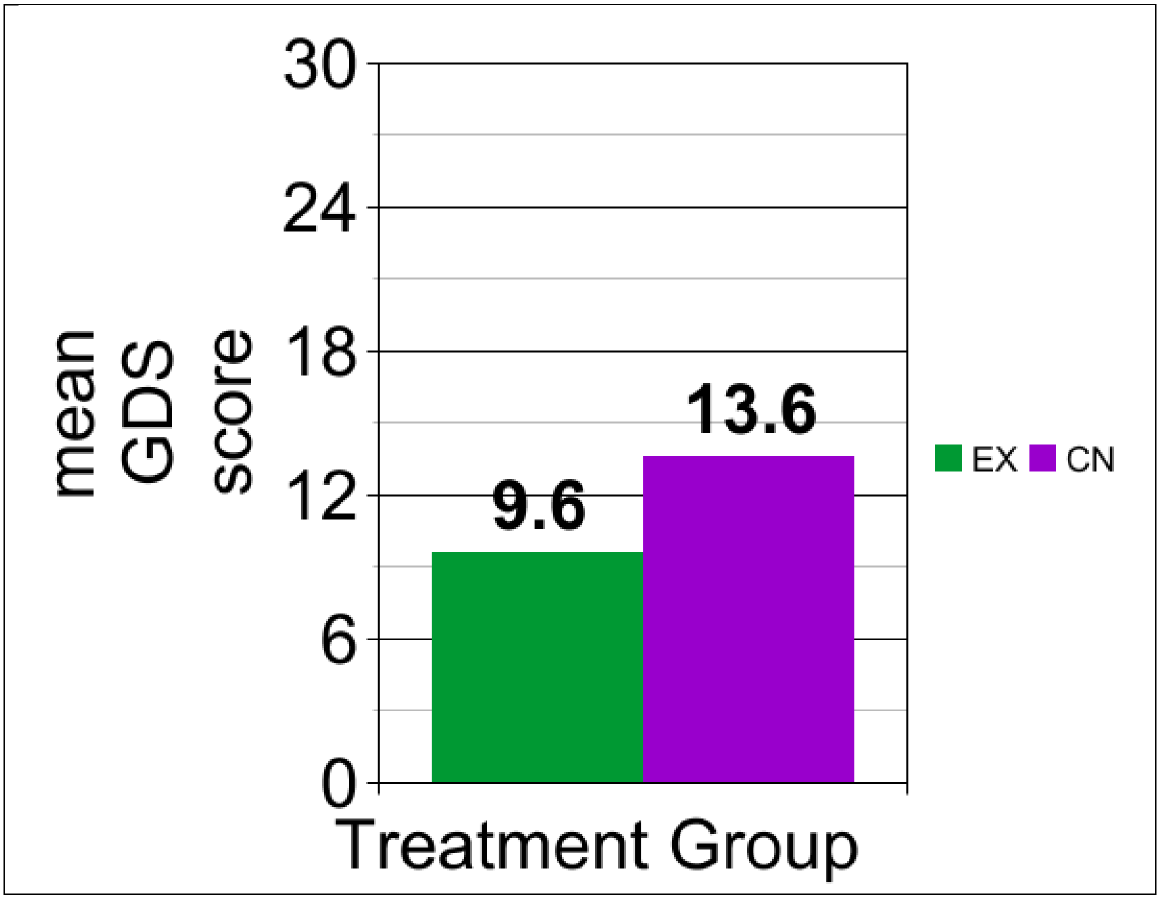
p = .002

difference between means = -4.0, 95% confidence interval

95% confidence interval of this difference: -6.483 to -1.517

Cohen's D = 0.81

The best type of chart to use would be a bar chart. An easy way to construct a colorful bar chart would be at this site.



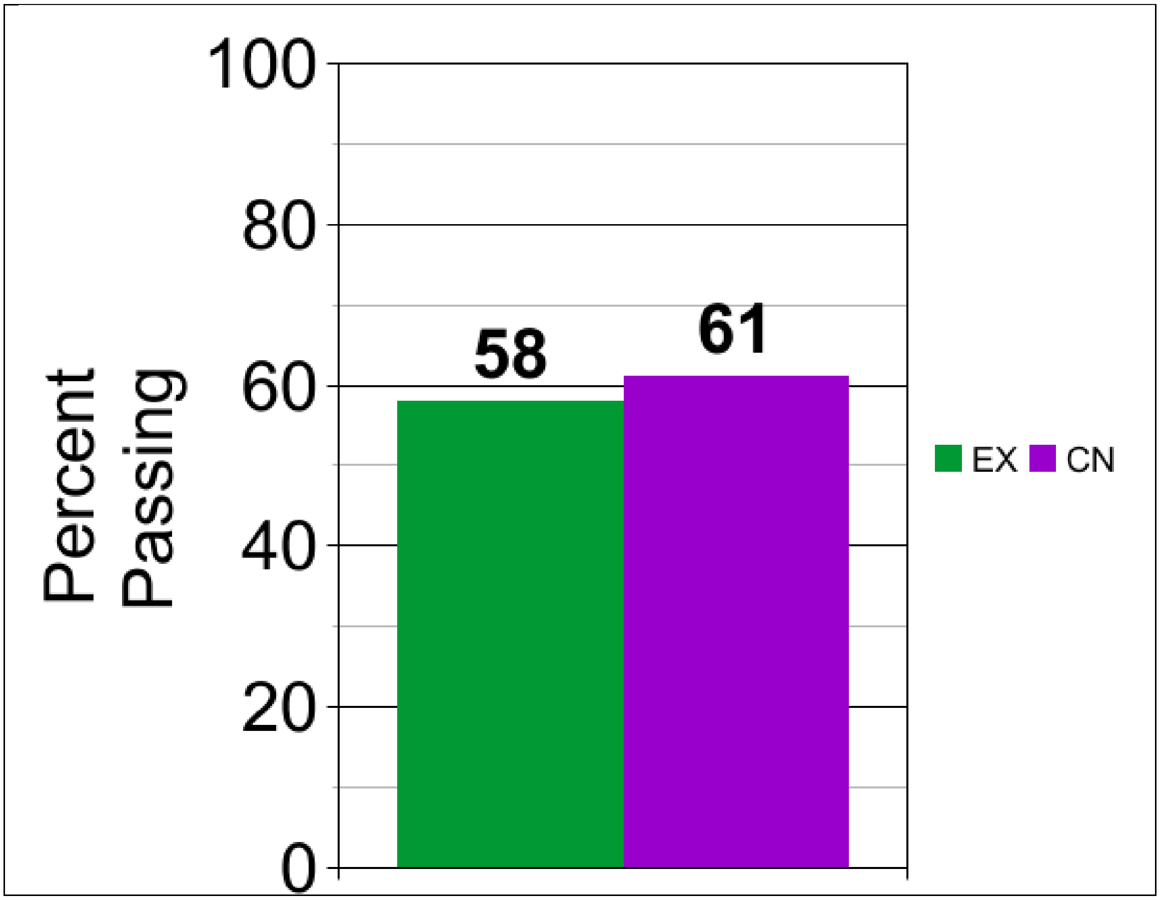
Another common situation is where we have a separate groups experiment, but the dependent variable is measured in a binary nominal scale (such as pass or fail). Let's take the example of a population of police recruits going through the academy. The topic is whether passing grades can be improved by providing tablet-based instruction. Let's assume that a sample of 30 is randomly assigned to the experimental group (n = 12) and a control group (n = 18). The outcome variable can be summarized on this chart.

Group	N	Number passing	Percent passing
Ex	12	7	58%
Cn	18	11	61%

p > .10

The inferential statistic used would be the Fisher Exact Test.

The bar chart is also the best graphic to us. Here the height of the bar represents percent passing as the central tendency of the dependent variable.



Another situation for a two group experiment is where the outcome variable is measured in a multiple nominal or ordinal scale (with perhaps four levels). Let’s take the example of the population of voters. Let’s see if a debate among the four leading candidates has an impact in terms of the outcome variable of preference. Let’s take a sample of 100. The experimental group (n = 40) is assigned to view an hour long debate between the four candidates, and the control group (n = 60) does not get to see the video.

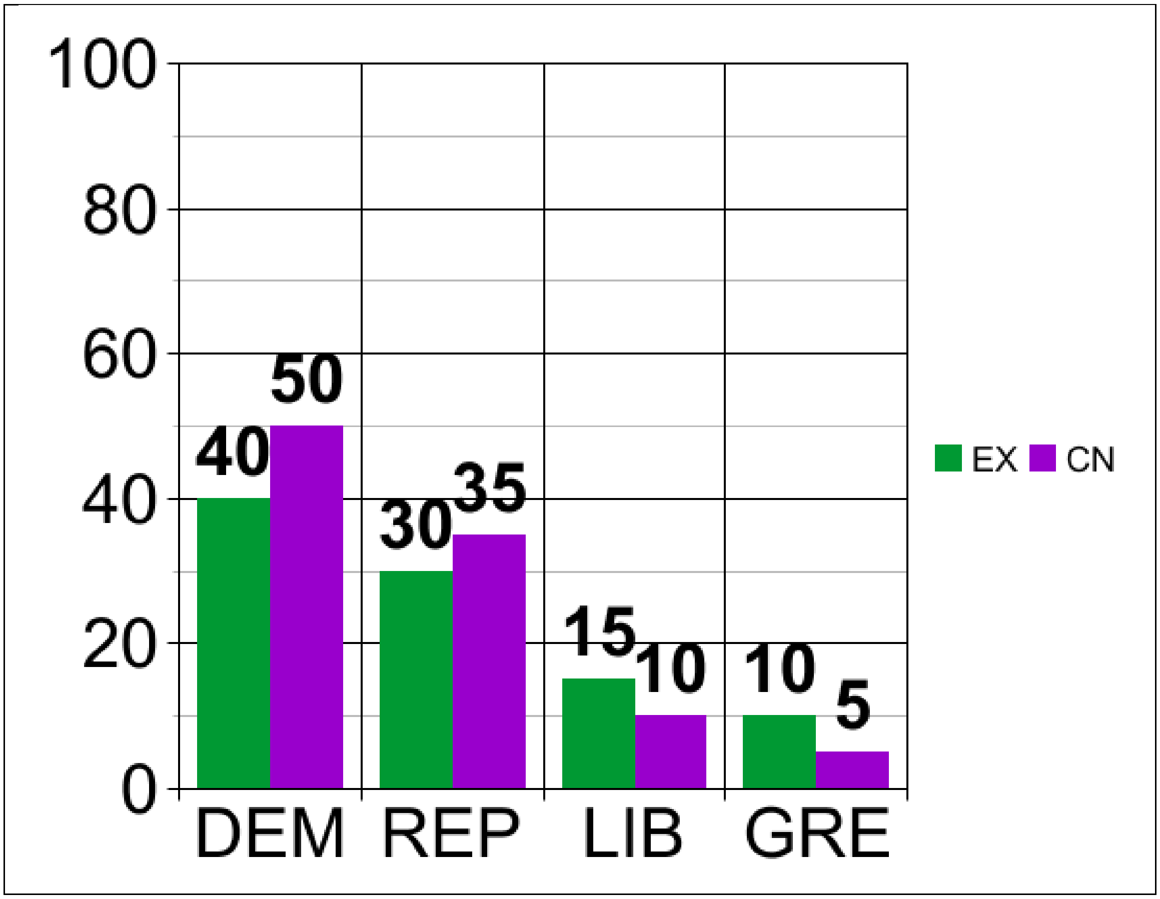
Group	N	Dem	Rep	Libertarian	Green
Ex	40	45%	30%	15%	10%
Cn	60	50%	35%	10%	5%

Kolmogorov-Smirnov p > .10

This site performs the two-group Kolmogorov-Smirnov calculation.

So, this difference between the groups is just not enough to be significant, given this sample size. Although there appears to be a trend for watching the video to be associated with increased support for the minor parties, the null hypothesis cannot be rejected.

Again, the bar chart would be the best graphic to employ. Now, it would be better to use four bars for each group, showing the distribution over that entire variable.



Final Thoughts on Experiments

Experiments remain the best way to verify that one variable has a causal relationship with another. The greatest limitation of experiments is that they are difficult to perform correctly: manipulating the IV just right, measuring the DV just right, controlling for confounding variables, and making sure that the group assignment was truly randomized.

Another limitation is that these two-group experiments are rarely sufficient to tell us how one variable caused another. Experiments are better at identifying causal outcomes than causal processes.

Chapter #10: Alternative Experimental Designs

Experiments are the best research method for inferring a causal relationship between the variables. So, why aren't experiments used all the time? Why is most of the research presented at conferences and published in peer-reviewed journals merely some form of a survey, measuring rather than manipulating the independent variables? The answer is that *between-subjects* experiments are hard to do (at least hard to do well). This chapter reviews some of the alternative designs: how to do something like experiments without random assignment to *separate groups*.

Quasi-Experiments

Probably the hardest thing to achieve is that of a truly *random assignment*. So, one of the most common approaches is to forget about the random assignment to experimental and control groups. It is more convenient just to take two existing groups and call one the experimental, giving it the treatment, and call the other group the control, and use it for comparison. Because this approach lacks random assignment, most scientists would say that it falls short on the criterion for being called a real between-subjects experiment (or at least a true randomized control trial, RCT). The term *quasi*-experiment is commonly used to describe such research manipulating an independent variable using convenient, existing groups.

Without randomization, the risk is that some *confounding variable* will arise. For example, suppose that instead of randomly assigning my student subjects to a treatment group (using the learning drills) and a control group (without the drills) I just said that one section of a course (MW 2) was going to get the drills, while the other section of the same course (TT 9) was not going to get the drills.

To the extent that these two classes are taught in exactly the same way, and have students with similar backgrounds, there may yet be *internal validity* to this quasi-experiment. But suppose that the MW 2 class is at the same time as the Organic Chem lab, and O-chem attracts the stellar, pre-med students. That would mean that the MW 2 class would attract few of these better students, but that time slot might dovetail nicely for the schedules of the students taking a remedial math class, so the MW 2 class would have a disproportionate amount of students who cannot understand percents, charts and tables.

The two groups being compared do not have to be of equal size, and this is true whether we are talking about a random assignment experiment or a convenient comparison of existing groups in a quasi-experiment. If you happen to have fifty subjects, and can assign them half and half so that exactly twenty-five are in the experimental group and twenty-five in the control group, that is OK, but because we will be comparing the groups in terms of central tendencies (e.g., percents, means, medians) it does not matter if group size is unequal. If you discover that one group is missing a questionnaire or two, you don’t have to remove an equal number of questionnaires from the other group. In general, the larger each group, the better.

Even if we have a quasi-experiment with somewhat unequal groupings, that should not be solved by artificially limiting sample size so that the two groups are closer in size. For example, suppose your experimental group was a class with only twelve students, but the control group has 52 students. If you can get data on all 52, use them. Don’t worry about other independent variables separating the two groups if you have those variables measured. Now, if you cannot measure those variables, maybe you could argue that it would be better to select a smaller sample out of the control group if you can make that smaller sample more comparable in background variables to the experimental group. (Also see matched pairs designs under repeated measures.)

When it comes to using tables and charts for quasi-experiments, just follow the suggestions that the previous chapter gave for separate groups experiments. Remember, the quasi-experiment differs from a true, randomized experiment in only way. It does not differ on design (separate groups) or on how the variable is measured (both the quasis and the randomized can use anything from binary nominal to continuous ratio scaling). The only difference is how the grouping takes pace: random assignment or pre-existing.

Case Control for Epidemiological Studies

Another interesting variation on separate groups designs is the *case control method* widely used in epidemiological studies. Technically, this is not an experiment or a quasi-experiment because the independent variable is simply measured rather than manipulated by an experimenter (but in the case of diseases, intentionally infecting a human subject may raise ethical questions). So, this is more of an epidemiological survey, usually just using a two-by-two contingency table, with the predictor variable being some event (usually exposure to something hypothesized to be the etiology of the disease) and the criterion variable being diagnostic confirmation that the patient has the disease. (Remember, we are using proper medical terminology, so *positive* means the presence of the disease, which happens to be a bad outcome, while *negative* means the absence of a disease, which is a good outcome.)

Contingency Table for Epidemiological Study

	DV = diagnosed positive	*DV = diagnosed negative*	*Totals*
IV = Exposure	A	B	A + B
IV = No exposure	C	D	C + D
Totals	A + C	B + D	N = A+B+C+D

A + B + C + D will represent our N, the sample size.

Let’s suppose this was an attempt to correlate cigarette smoking (the exposure) and lung cancer (the resulting diagnosis).

The marginal A + B represents all those smokers in the sample.

The marginal C + D represents all those non-smokers in the sample.

The marginal A + C represents all those in the entire sample eventually diagnosed with lung cancer.

The marginal B + D represents all those in the entire sample who were never diagnosed with lung cancer.

Cell A represents the smokers who were diagnosed with lung cancer.

Cell B represents the smokers who were not diagnosed with lung cancer.

Cell C represents the non-smokers who were diagnosed with lung cancer.

Cell D represents the non-smokers who were not diagnosed with lung cancer (probably the largest cell since most people don’t smoke and most people don’t get lung cancer.

Most of the time the case control method is used, it is *retrospective*. In other words, it starts with the present diagnosis as the grouping variable. Suppose we start with a group of 50 lung cancer patients and a comparison group of 50 subjects who have never been diagnosed with lung cancer. (We could try to control for confounding variables by making the two outcome groups proportional in terms of background variables like gender, ethnicity, age, education and socio-economic status. We then go backwards in time and see how many of each outcome group were smokers, enabling us to fill in cells A, B, C and D. Except for the control of background variables, the case control method is not much better than a *correlational* survey where we look into a hundred patient charts and ask 1) is this patient a smoker, and 2) did this patient develop lung cancer.

When it comes to using statistics, tables and charts for case control designs, we can stick with percents (as our descriptive), either Fisher Exact or Yates Chi Squared (for the inferential) and bar charts (for the infographic). Correlation coefficients could be used (especially as part of a correlation matrix) where we can quickly view the relative level of association between the outcome variable and several predictors.

However, instead of correlation coefficients, many epidemiological investigators prefer to use the *odds ratio* with the case control method. What are the odds that a lung cancer patient was a smoker? What are the odds that someone who does not have lung cancer was a smoker? Divide the first odds by the second. That odds ratio tells us how much more likely lung cancer patients are smokers. An odds ratio of 1.00 (like a correlation coefficient of 0.00) states that there is no relationship between exposure and outcome. The higher the odds ratio, the greater the association. Odds ratios are calculated at this MedCalc site.

Compared to a correlation coefficient that might be derived from the two-by-two contingency table (e.g., phi) the odds ratio really picks up on small differences, and is frequently used when a disease is fairly rare.

Another frequently used measure of association with this design is *relative risk.* This is calculated by finding the proportion of those exposed who develop the disease and the proportion of those not exposed who develop the disease (and then dividing these proportions). Once again, 1.00 means no relationship, while the higher the relative risk ratio, the greater the relationship between exposure and outcome. Here is a calculator for relative risk.

Sample vs. Norms

Another alternative to the separate groups design for an experiment is to do sample vs. norms, but this approach usually introduces many confounding variables. Suppose the dependent variable is attitudes about the President's handling of foreign policy. We have national norms on that, thanks to a Gallup Poll. So, I get my sample of student volunteers from Crafton Hills College, and the experimental treatment I give them is to listen to some speeches by foreign policy experts praising the President's stand. I then measure the students' attitude about the President's handling of foreign policy, and if it is more favorable than the Gallup Poll's norms, I conclude that those speeches were convincing.

Confounding variables would include the fact that all of my subjects were students, all were from southern California, there was a month's time delay between when Gallup did the Poll and I measured my students' attitudes, and perhaps only people with greater political interests volunteered to be in this experiment. Any of these variables (rather than the speeches) might account for the difference between my sample and the national norms. Indeed, some of these factors might predispose my sample to be more favorable to the President, and some could reduce favorability, and I have no way of knowing the direction, let alone the strength of any of these impacts (unless we also do a correlational survey with the data, but that would not get around the problem of spurious correlations).

Another problem especially heightened by sample vs. population experiments is the *Hawthorne* effect. This sample was singled out for special treatment, and the rest of the population was not. On many measures of performance and attitude, the presence of the researcher or even the subject's knowledge that he/she is being observed influences the subject's behavior. Subjects might perform better for an audience, or act in more socially acceptable ways in order to get approval. This is less of a problem in separate groups experiments, because both groups know that they are being observed. We have a real problem, however, when the comparison data come from population norms that were based upon organisms who did not think that they were being so carefully studied.

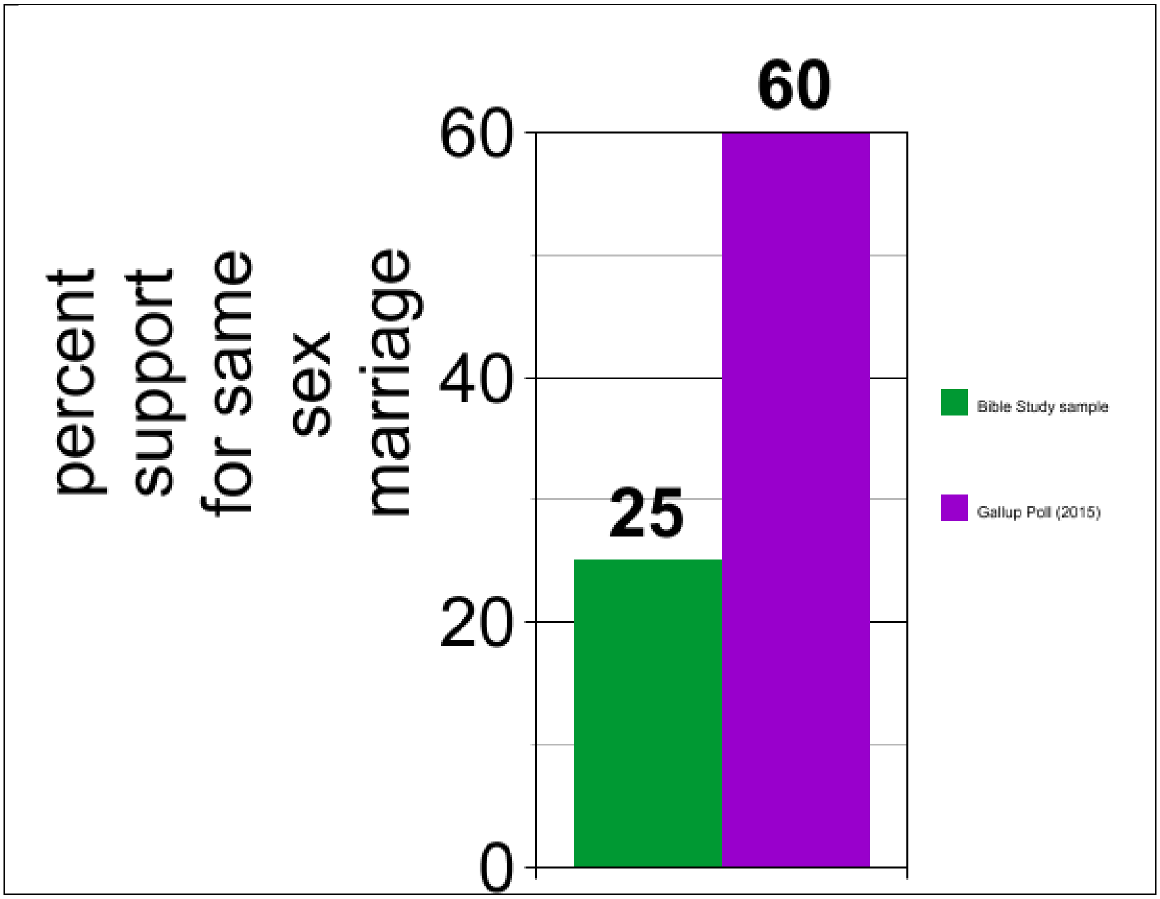
For example, let's look at the cognitive performance of geriatric patients during the first year after a diagnosis of dementia. We know that the natural course of this disorder is a gradual deterioration. Now, imagine that a small sample has been selected (even randomly selected) to receive a new form of memory training based upon computer simulations. Suppose that the resulting data show that the treated sample deteriorated less (significantly less) during that first year after diagnosis. The difference in performance might be related to the new memory training, or just to the fact that the treated sample got more human interaction with their trainers and evaluators. A better design for this study would have been separate groups, with random assignment, with the control group receiving a placebo. In this situation, an appropriate placebo might take the form of an equal number of contact hours with trainers and evaluators who did something else with those elders (instead of the computerized memory training given to the experimental group).

Tables & Charts for Sample vs. Norms

In the case of a variable scored on a ratio or interval scale, we can compare sample and norms using a mean or median. Suppose that we want to know if meditation increases life satisfaction (as measured by the Cantril Ladder). So, we get a sample of student volunteers (N = 20) and have them practice mindfulness meditation every day for a week, then measure where they are on the ladder of life satisfaction. For descriptive statistics, we could compare the sample mean or median to the national norms (i.e., the mean from a recent Gallup Poll using the ladder). Suppose there was a national median of 7. We could also describe our sample in terms of the percent of the sample above the norm (the national median). This directly compares the sample and the population, because by definition, 50% of the population is above the median, and we could use the Binomial Distribution to test for significance. Just enter the sample size as the number of trials, and enter the number in the sample who were above the median where it asks for successes.

What should be most obvious in this example is the great number of confounding variables. Rather than actually employ this design, it would be better to do a randomized separate groups (described in the last chapter) or even a quasi-experiment, described in the previous section of this chapter, or even a repeated measures design, described the next section of this chapter.

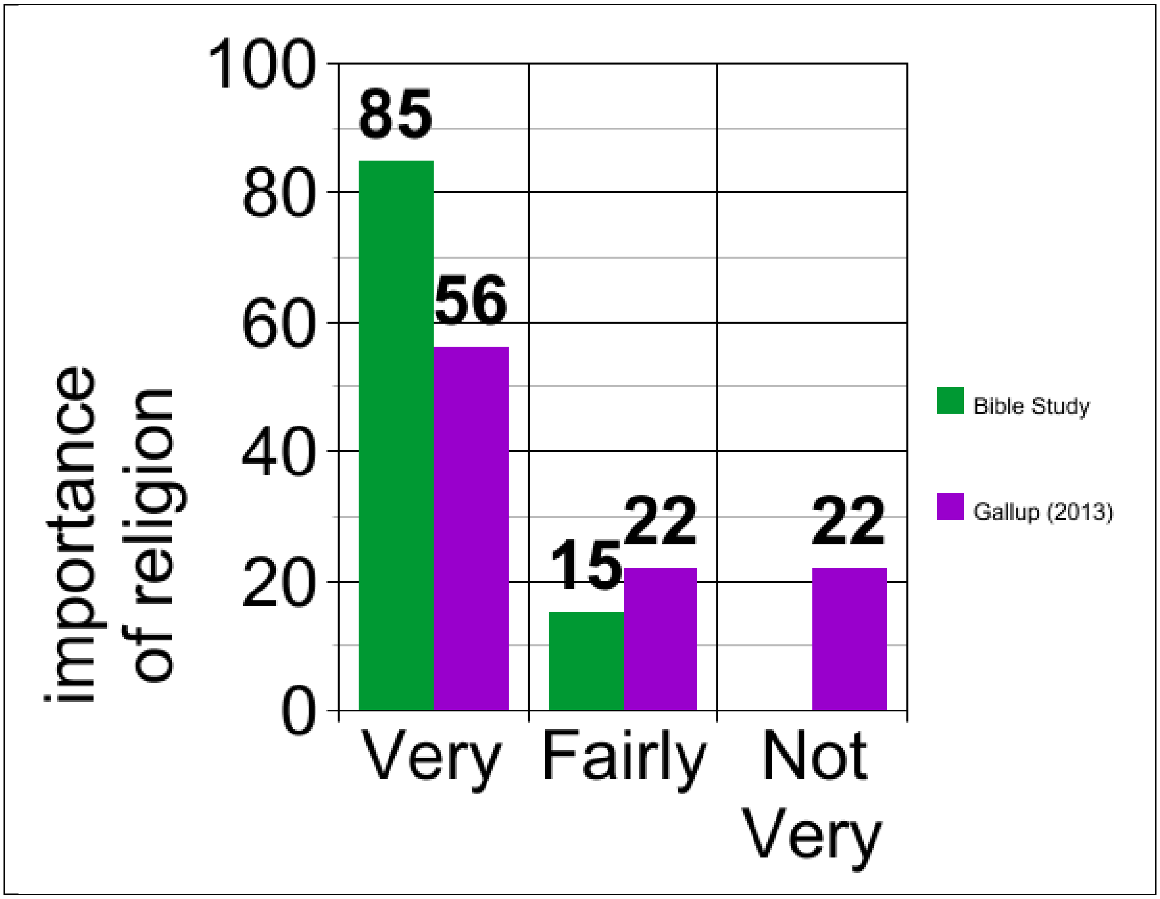
Where the criterion variable is measured on a dichotomous scale, we are just going to compare percents. Suppose we want to see if a sample of Bible study participants at a local evangelical church differ from the national norms in terms of attitudes about marriage equality for LGBT. Suppose that you see a national figure of 60% approval of marriage equality, published in a Gallup Poll. However, your Bible study sample (N = 32) shows approval by only 8 (in other words, 25%). Again, the binomial distribution can test for significance. (Use the national norm of .60 instead of the default .50), enter sample size of 32 where it asks for number of trials and 8 for successes. Bar chart is the best way to depict the results visually.



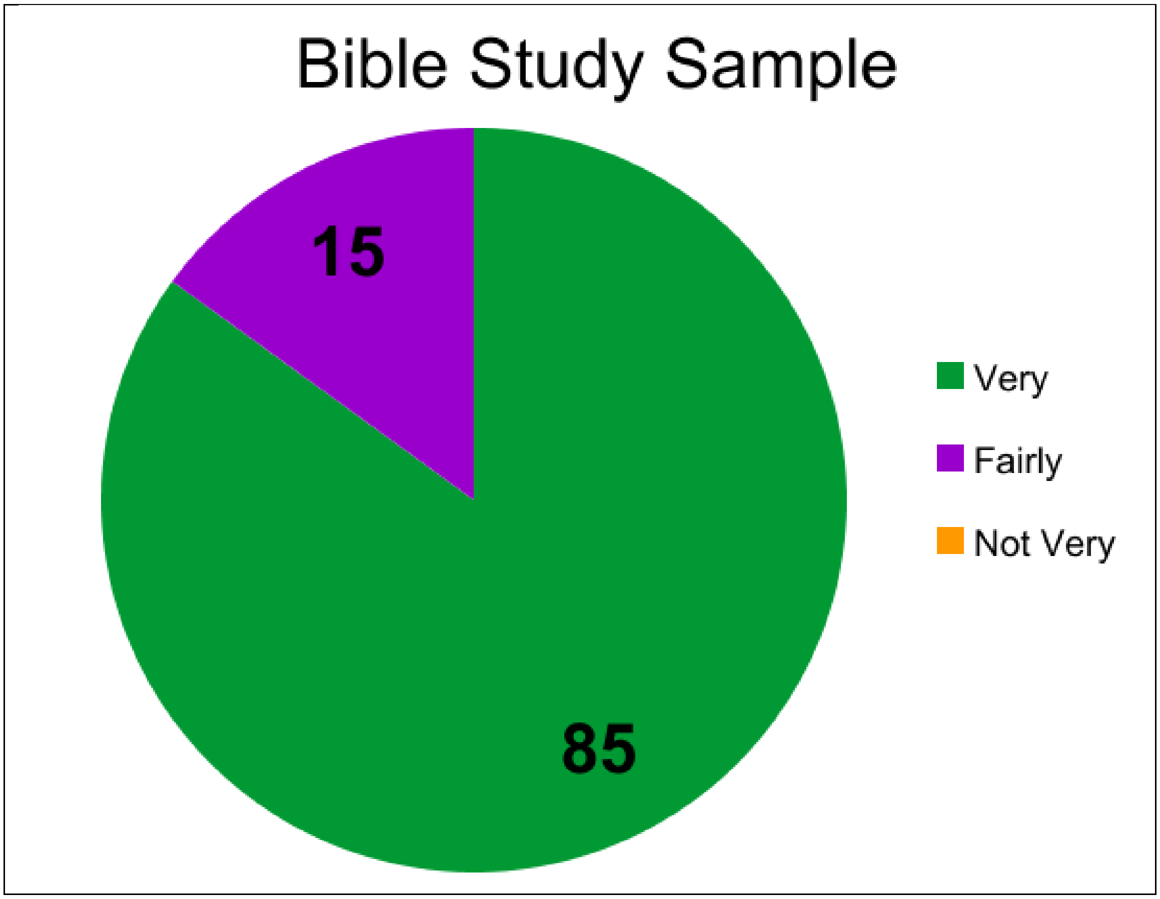
When the criterion variable is measured in a multiple nominal scale (or an ordinal scale with several levels) a more comprehensive table and bar chart should be used to visually depict the data. Let's take that same sample (the Bible study participants) and look at a different variable: religiosity. There are Gallup norms for religiosity, measured on a three level ordinal scale: "How important is religion in your life: very important, fairly important, not very important."

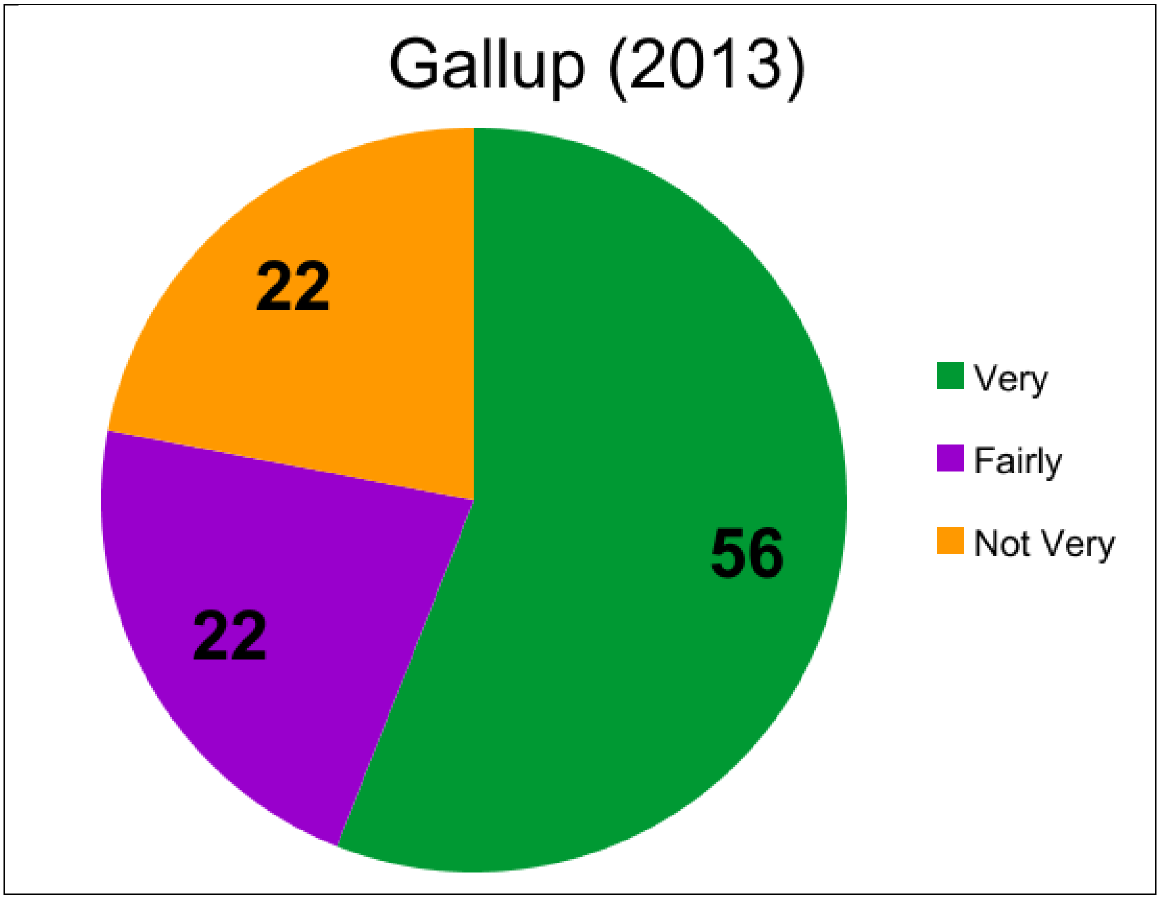
Religiosity	Very Important	Fairly Important	Not Very Important
Bible Study	85%	15%	0%
Gallup (2013)	56%	22%	22%

The inferential statistic would be the Kolmogorov-Smirnov test for absolute maximum difference between cumulative distributions of sample vs. population (p < .001 for the data in the above table). The bar chart should show the breakdown across the three levels for both the sample and the population norms.



I usually don't like to call on pie charts when doing a comparison, but in this case of extreme ceiling effect in the sample, the difference between the two pie charts makes the illustration vivid.


Bible Study Sample
15
85
Very
Fairly
Not Very



Repeated Measures

A superior, and possibly easier, alternative to the between-subjects experiment is often a *within*-subjects experiment. *Repeated measures* designs have many advantages (for both experiments and surveys). The inferential statistics can attain significance with a smaller sample size. Instead of using randomly assigned separate groups, a within-subjects experiment uses repeated measures of the sample. Instead of hoping that confounding variables are randomized when we put different people into experimental and control groups, the within-subjects experiment sets each participant as his/her own control. So, we end up with experimental and control conditions, rather than experimental and control groups.

Not every repeated measures design is an experiment. If we just measure the same variable twice, without any real manipulation of a stimulus between the time that the different measurements of the same variable were made, this would not meet our essential criterion to be called an experiment.

In order to achieve these advantages of repeated measures designs, the data have to be carefully coded so that we can match each subject's first measure with that same person's subsequent measure. It is easy if all the data are obtained from the same paper questionnaire (or archival record) and we just enter these numbers in separate columns on a spread sheet, one column for the first measure of the dependent variable and another column for the second measure of the dependent variable.

In the example below, imagine that each subject filled out one questionnaire asking about background variables like gender, age, ethnicity, and current marital status. Then each participant expresses his or her attitude about Toyota automobiles by responding to a statement "Toyotas are well made automobiles." The subject is given a five-level Likert scale for responding. Then, each subject watches a video about a news report on faulty air-bags. The participants are given another opportunity to express their attitudes about Toyota quality.

MALE	AGE	HISP	MARRIED	BEFORE	AFTER
1 = yes	1 = over 25	1 = yes	1 = yes	5 = strongly agree	5 = strongly agree
0 = no	0 = under 25	0 = no	0 = no	4 = mostly agree	4 = mostly agree
				3 = not sure	3 = not sure
1 = male			1 = married	2 = mostly disagree	2 = mostly disagree
0 = female			1 = widowed	1 = strongly disagree	1 = strongly disagree
			1 = divorced		
			1 = separated		
			0 = never married		
0	0	1	0	3	2
1	0	1	0	4	2
1	0	0	0	4	1
1	0	0	1	5	3
1	1	1	0	2	2
0	1	1	0	4	3

It is very easy to transfer the data from a single questionnaire sheet containing all this information (i.e., both the before and the after ratings of Toyota) on to a multi-column spread sheet. But suppose you feared that some subjects might be tempted to go back and change their initial (before) ratings after they had seen the news story, and so you collected the initial questionnaires and then gave out new questionnaires after the news story was shown. How do you now match up the fifth subject's before answer that same person's after answer? If you use identifiers like names or student ID numbers, that raises the risk of losing the subjects' anonymity. You need to come up with some other secret coding that allows you to keep each subject's data together without compromising the subject's identity.

Go back to the example of using drills in my classes to improve student performance. A repeated measures design would have students in the sample prepare for the first quiz without the drills, and prepare for the second quiz with the drills. In each case, we could see if that particular student did better, worse, or the same with the drills. This would control for factors such as gender, ethnicity, age, motivation, previous coursework, and other things in the students' background. Since I, as the instructor, already know the identity of each student getting each quiz score, subject identities are not further compromised by a loss of anonymity (I just have to respect confidentiality in my reporting of the data so that no outside person could know how Sara Garcia scored on a particular quiz).

Unfortunately, some new confounding variables may come into play with repeated measures. Are the two measures really of the same performance (or was test #2 just easier than test #1)? One way to get around that would be to *counterbalance*: have half of the students use the drills for unit #1 and the other half use the drills for unit #2. When we do such counterbalancing, we are also introducing a separate groups design: those who received one order of the questions vs. the other order of the questions, so you can develop hypotheses that can be tested by that separate groups design.

We also have to consider if the research involved any *carryover* effects, such that the process of measuring the first time may have influenced subsequent measures. *Practice* effect that might boost scores the second time around. Suppose we are doing a study on video games. We tell gamers to play a new video game with the left eye closed, then we measure their performance again, this time with the right eye closed. The second time they play the game, they have more practice, and they may do better regardless of which eye was being used.

Sometimes we have the opposite impact: *fatigue* (or boredom) reducing performance the second time around. Do students run faster in Nikes or New Balance running shoes? OK, everybody put on your Nikes and run a mile so I can record your times. Now everybody, change shoes and run around the track again so I see how fast you are in the New Balance shoes. If everybody now gets a slower time is it because of the different shoes or because the runners are still tired from the first run? If so, Nike has the unfair advantage of having the subjects when they had more energy. *Boredom* with a task could also reduce performance on subsequent retests. Counterbalancing (have half use Nikes the first time, the other half use New Balance) can reduce these carryover effects.

Priming comes into play just by getting our subjects thinking about a topic before we ask the dependent variable question. Specifically, asking the first measure may influence the answer on the second. Asking subjects how upset they are by heroin overdose deaths and then asking them about firearm deaths may suppress the second variable ("How can I be more upset about fewer deaths"?). Priming can also be a problem in any design, not just repeated measures, because a completely different question may serve to prime later answers about later variables. For example, asking subjects' GPA may prime later answers about self-assessed cognitive abilities. Most priming effects can be dealt with by counterbalancing.

Confounding Variables with Repeated Measures	
	Impact on later measure
Carryover effects	
Increasing motivation	Increases performance
Boredom	Decreases performance
Practice	Increases performance
Fatigue	Decreases performance
Priming	??
Natural course	
Disease prognosis	Improves (e.g., depression) Deteriorates (e.g., dementia)
Natural improvement	Getting better at a job
Age related	Improves (life satisfaction) Deteriorates (physical)
Attrition	Loss of subjects due to lifestyle (death) Or poor performance (discharging employees)
Cohort related	Attitudes due to historical context

Then we have to consider an increasing (or decreasing) trend independent of treatment. In industrial psychology, most newly hired workers' performance on the job will improve over time, whether or not they receive additional training or incentives. For clinical studies, was there a natural course to the disorder? We know that the common cold's symptoms tend to improve over ten days, regardless of treatment. Many cases of depression spontaneously improve over six weeks (while fewer worsen). Almost all cases of Alzheimer dementia worsen over a year, regardless of treatment. (A treatment is considered effective if it is able to slow the deterioration.)

Another problem with a simple pre-test / post-test clinical study would be the Hawthorne and expectation effects. The sample might improve over the study, but could that be due to their increased motivation and hopes being raised? There is no placebo control group to compare with to control for these confounding variables.

Many surveys on lifespan development use a repeated measures design, involving an *interrupted time series*. It is known as a *longitudinal* study, and follows a given sample (all from the same age *cohort*) over a long period of time. This has some advantages over a separate groups survey (known as *cross-sectional*) in which different cohorts would be compared. For example, suppose we want to study attitudes about marriage and see if they change over the lifespan. So, we do a cross-sectional study: get a group of 20 year olds and another group of 80 year olds and ask about attitudes toward marriage. But if we find that the 80 year olds have more traditional attitudes about marriage, is that because they are 80 years old or because they were born before World War II? Can we expect that our current 20 year olds, born after the development of the internet, will also develop those same traditional attitudes by age 80? Cross-sectional research cannot answer that question.

The best (but most difficult) longitudinal studies start in the present time and then wait for years until the study is over. In some of the most famous *prospective* studies (e.g., the Terman study of gifted children) the original investigators passed away and subsequent generations of researchers had to continue on. However, the longitudinal design also brings a major problem: subject *attrition*. Some of the 20 year olds won't be around to be re-tested at age 80. Unfortunately, those who die off (or otherwise disappear from our sample) will disproportionately represent those who are gay men, drug users, less religious and unmarried. We may end up with a repeat sample over-representing Catholic nuns, Seventh-day Adventists and Mormons, and they are going to express more traditional views of marriage.

Another longitudinal approach is *retrospective*. We measure where the subjects are currently in their present lives and then have them remember where they were on the same variable in the past. For example, Gallup can use the Cantril ladder to ask subjects how they are today, and where they stood on that same ladder of life satisfaction five years ago. Validity of the past measures can be affected by a deteriorating recall (or the tendency of humans to rewrite their own past in order to correspond with present needs). Similarly, a longitudinal study could be hypothetically prospective, comparing where the subjects are today on a variable, and also asking them (today) where they predict they will be at some point in the future. For example, we could ask on which rung of Cantril's ladder the subject thinks he will be standing in five years.

Another use of repeated measures designs comes when we want to look at subjects' attitudes on different aspects (or performance in different areas). For example, we can ask a sample of workers about their level of satisfaction with different aspects of their jobs: pay, benefits, work schedule, supervision, opportunity for advancement. We can see which of these aspects has the greatest satisfaction (and the least). We can present different scenarios and have the subjects respond to each: "Is spanking an appropriate punishment for a five-year-old boy who …? what about for a ten year old boy who …?" Of course, when we ask this series of questions about related aspects, there is the chance that the order of those questions might influence the answers. If this is likely, we can use counterbalancing.

It is not a repeated measures design to give subjects a *concurrent* forced choice between two alternatives: "Which is the better truck: Ford or Chevrolet?" A repeated measures design would ask two questions, one about each alternative, and get a rating on each, and then compare the two. Let's understand this difference between concurrent and repeated measures in terms of how we code things on a spreadsheet. A real repeated measures design would have two columns, one for the rating of Chevy and another for the rating of Ford, while the concurrent design would have one column for preferred truck, scored Ford = 1, Chevy = 0.

Notice how much more precise the repeated measures would be. It would give us everything the concurrent forced choice would give us (whether the subject gave a higher rating to Ford or Chevy) plus it allows us to see cases of a tie or no preference, where Ford and Chevy scores the same. An additional advantage is that we could also use these individual ratings for Ford and Chevy to do a correlational design with background variables (e.g., gender, age, ethnicity).

One way to improve a concurrent measure's precision would be to move from a binary nominal to an ordinal scale. "How much would you prefer one of these trucks: Ford or Chevy?"

STRONGLY CHEVY

SLIGHTLY CHEVY

NO PREFERENCE

SLIGHTLY FORD

STRONGLY FORD

We could label this variable Chevy preference and code it 5,4,3,2,1 depending upon the level of the response.

We could get something similar if we started out with two columns, one for Ford and one for Chevy. We could then create a composite variable (Ford over Chevy) in yet another column that looked at the difference between the two. So, a subject who gave Ford a 4 and Chevy a 3, would get +1 in the difference column, while someone who gave Ford a 3 and Chevy a 5 would get –2.

Another use of the repeated measures design is when we have data from *matched pairs*. For example, suppose we ask heterosexual married couples (n = 50) their level of satisfaction with the marriage: extremely, very, somewhat, slightly, not at all. We could then use a repeated measures design to compare each husband's level of marital satisfaction to that of his own wife's level in order to test the hypothesis that men are more satisfied with marriage than women are.

In these matched pair designs, we change one rule about coding on a spreadsheet. We usually insist that each subject be on one row and everything on that row be about only that one subject. However, with matched pair designs the statistical unit is the pair, not the individual person. So, in the above example of fifty married couples, the sample size is 50, not the hundred persons (50 wives + 50 husbands). Each column would have information about the couple, such as years married, number of children, age of husband, age of wife, husband’s reported level of satisfaction and wife’s reported level of satisfaction.

When we have repeated measures data, we can also do a correlation between these measures. That will answer a different set of questions. The repeated measures design will answer if the subjects scored higher on the second measure or the first. The correlational design will answer if the same subjects who scored higher on the first measure were the ones who scored higher on the second measure. In the above example of marital satisfaction, the correlational design will answer the question if individual husbands and wives agree on the quality of their marriage: if a given husband says that he is satisfied, will that man’s wife also say that she is satisfied? A positive correlation would be hypothesized if we assume that one spouse’s happiness (or sadness) in a marriage is contagious: the mood will spread to the spouse. A negative correlation would be hypothesized if we assume that each marriage is a zero-sum game, and that if the wife is winning the arguments, she is elevating her happiness as the expense of her husband (or vice versa).

In order to use the repeated measures design, we have to measure the same variable more than once for each subject. If we measure a child's IQ at age six and then that child's SAT at age 18, that is a correlation of two different variables, not a repeated measures study of the same variable twice. It is true that the two variables, IQ and SAT, were measured at different times, but their means cannot be compared (especially in this example, where there are entirely different scoring ranges). You could correlate those two variables (and that would tell if the same children who scored higher on one variable at age six were also the ones who scored higher on the other variable at age 18), but such data could not tell us if the children had a higher IQ at age 18 (because we did not measure their IQs at age 18) or if they had higher SATs at age 18 (because we did not measure their SATs at age 5).

In order to use the repeated measures design, we have to be able to code our data as follows: all the data for each subject (or coding unit, when we have matched pairs) must be entered in the same row. If I am measuring workers' (n = 32) performance before and after training, I have to know that Jones scored 75 before and 87 afterward. Maybe I don't need to know that those two scores belong to Jones (we can preserve his anonymity) but I need to enter both scores from Jones on the same row (e.g., line 26) of spreadsheet. If all I have is 32 questionnaires filled out by workers before training and another 32 questionnaires filled out by the same workers after training, and I cannot match Jones' before to Jones' after, I cannot put his data in the same row on Excel, and I cannot use repeated measures statistics (or correlate the before scores with the after scores). I could treat the data as if they were separate groups (a before group and an after group), but I have introduced all the problems of a within-subjects design (e.g., attrition, practice effect), with none of the benefits (control of background variables).

Now let’s look at another previous example, the fifty married couples. Even though we have a hundred people filling out the questionnaires, the sample size is 50 (couples). We will need only 50 rows on Excel to enter the data, because both the husband's evaluation and the wife's evaluation will be on the same row. If we have not coded the questionnaires such that Mr. Green's data can be paired with Mrs. Green's data, and Mr. Brown's data can be paired with Mrs. Brown’s data, then we have to treat these questionnaires as if they were separate, unrelated groups of men and women.

Perhaps the worst of all designs is where some (but not all) of our subjects get both levels of treatment (and we don’t know which subjects those are). This is worse than a repeated measures design because we cannot match up the before and after scores for each individual, and worse than a separate groups design because we don’t know which subjects in the control group may have also been exposed to the experimental treatment. For example, suppose I want to do a separate groups quasi-experiment on the effectiveness of a Hillary Clinton campaign ad. I select my on-campus MW 10 General Psychology class as the experimental group, and show them the ad (independent variable) and then measure the attitudes about Hillary Clinton (dependent variable). I then get my buddy, Professor Cervantes, who teaches the MW 1 Philosophy class to use his students as the control group. They don’t get to see the Hillary Clinton ad (independent variable) but they also fill out a questionnaire measuring their attitudes about Hillary Clinton (dependent variable). Just try to list all the potentially confounding variables. Maybe psychology attracts more Democrats than philosophy does. Maybe the morning class attracts more single mothers. Maybe the philosophy class has discussed the ethics of abortion. Maybe students of a certain ethnicity are more attracted to one professor’s classes. But there is even a bigger confounding variable. Perhaps some of the students in my General Psyc class are also in Cervantes’ philosophy class, and saw the advertisement earlier that morning.

If we are having a separate groups design, each subject should be in just one group or the other, not in both. If we are having a repeated measures design, each subject should receive both measures, not just one (and we have to match each student’s before score to that same student’s after score). One solution in this example about the psyc and philosophy classes would be to say “If you already took this questionnaire in Professor Brink’s psychology class, don’t do it again here.”

Many designs you might imagine to be repeated measures don’t really qualify and cannot use the special inferential statistics comparing each subject’s before scores with that same subject’s after scores. Suppose your population is

- Patients in an acute care hospital in June and then a month later in July (most of the June patients have gotten better and been discharged, and some have died, and most of the beds in that hospital are now occupied by different patients)

- Undergraduate students at the University of Redlands in 2010 and again at 2015 (most of the students back in 2010 have graduated, some dropped out or transferred, and you would be lucky to find a handful who are still enrolled at the U of R working on the same degree)

- Employees at Walmart today and ten years from now (most of the workers might stick around for the next decade building up seniority, but many will find employment elsewhere, some will retire, drop out of the labor force, die or get fired).

- Customer satisfaction at an internet bundle provider last year and this year (but a new competitor has come into the local market with cut-rate connections, taking most of the price-sensitive customers away, just leaving those who appreciate this company’s quality service, even though it is expensive)

If you are in a situation like these, consider this as a design comparing shifting aggregates rather than repeated measures of the same subjects. (More about this in a later unit.) Use a time series line graph to indicate your comparison of the different time periods.

Another constraint on all repeated measures designs is that we cannot have any difference in how the criterion variable is measured each time. Clearly, we would not allow this situation in a separate groups design: with the experimental group’s performance being tested one way and the control group’s performance being tested another way. Unfortunately, as time goes on, measures of variables change. Suppose a factory made smaller widgets a few years ago, but now has shifted to large ones (which take a little longer to produce). In the meantime, we have given our workers some training, and need to see if it has worked. So, we compare their pre-training productivity with their post-training productivity, but the measurements are not the same (and this has even added a confounding variable). By changing the measure of productivity, we end up comparing apples and oranges.

Or suppose that before training, we measured worker performance on a four-level ordinal scale: excellent, good, fair, poor. Now suppose by the time training is completed we have changed the wording of some of the levels, or come up with a new five-level ordinal scale: outstanding, commendable, average, barely acceptable, unacceptable. We might use our accumulated data to suggest which scale might be better, but we cannot use these data to tell us if the training that happened in between these measurements caused worker performance to improve or deteriorate.

Sometimes the problem is even worse, and we don't even have the same operational definition of the variable being measured during the two times. Consider a matched pair design of husbands and wives, and we are trying to see if marital satisfaction is greater among men or women. We use a dichotomous (yes/no) response format. We ask the husbands: "Do you still want to stay married to your wife?" We measure marital satisfaction in a slightly different way with the wives: "If you had to do it over again, would you have married the same man?" We could take either yes answer as implying marital satisfaction, but if the percentages differ, does that mean that men have more marital satisfaction than women do, or does it mean that it is easier to get a yes answer from the first question than from the second?

If we have coded the repeated measures data correctly and can enter them all on the same row for the same subject (or matched pair), then we can do correlations by selecting a Pearson (or Spearman) for *correlating* the different columns and we select a t test (for paired data, also known as dependent samples) or a sign test for *comparing* the central tendencies of the different columns. The correlational design answers one question: if one member of the couple is satisfied does the spouse also tend to be satisfied? The comparison of the two columns answers a different question: are wives more satisfied with their marriages than their husbands are?

If we just have two repeated measures (e.g., before and after, husband and wife) and one criterion variable measured in a binary nominal scale (e.g., pass/fail, yes/no) we could use the old, less sophisticated two-by-two contingency table to tabulate and report our data.

Most of the time students try to put repeated measures data into a contingency table, they get it wrong, because they just take the type of table they would construct for a separate groups design and plug in the repeated measures.

Contingency Table for **Separate Groups** (correct)			
	DV = yes	*DV = no*	*Totals*
IV = yes *(group one)*	A	B	A + B
IV = no *(group two)*	C	D	C + D
Totals	A + C	B + D	N = A+B+C+D

The proper inferential statistical test to use with the above table would be the Fisher Exact Probability or the Yates version of Chi Squared.

Contingency Table for **Repeated Measures (incorrect)**			
	DV = yes	*DV = no*	*Totals*
IV = yes *(measure one)*	A	B	N = A + B
IV = no *(measure two)*	C	D	N = C + D
Totals	A + C	B + D	2N = A+B+C+D

The problem comes with sample size and the use of proper inferential statistics. The above table mistakenly claims that we have double the sample size just because we asked our sample the same question twice. If we were to use the Fisher Exact test or Yates version of Chi Squared for our inferential statistic, we would be misled into thinking that statistical significance was better than it actually is.

Here is the rule of thumb for contingency tables: each subject should appear in one and only one cell: A, B, C, or D. Here is how to arrange our data correctly for a repeated measures design with a dichotomous variable.

Contingency Table for **Repeated Measures (correct)**			
Second Measure			
	DV = yes	*DV = no*	*Totals*
First measure *DV = yes*	A	B	*A + B*
First Measure *DV = no*	C	D	C + D
Totals	A + C	B + D	N = A+B+C+D

Cell A represents those subjects who answered yes both times.

Cell B represents those subjects who answered yes the first time, but switched to no the second time.

Cell C represents those who answered no the first time but switched to yes the second time.

Cell D represents those who answered no both times.

The marginal A + B represents all those who answered yes the first time.

The marginal C + D represents all those who answered no the first time.

The marginal A + C represents all those who answered yes the second time.

The marginal B + D represents all those who answered no the second time.

Now, A + B + C + D will represent our N, the sample size.

Subjects in cells A and D did not change, so the independent variable manipulation had no impact on them. If cells B and C are approximately equal, then there was no clear pattern of change (from yes to no, or from no to yes). Only if there is a great difference between the number in B and the number in C can we say that the independent variable helped (or hurt).

The appropriate test statistic would be the Binomial Distribution, Here your number of trials would be B + C, and the number of observed "successes" could be either B or C (don't worry which one because the two-tailed test gives the same answer). Leave the probability at 0.5.

Another option would be to use the McNemar version of Chi Squared.

Tables & Charts for Repeated Measures

Here are some tables and charts we might use with repeated measures. (They are pretty much like what we saw for separate groups experiments, except that now we are comparing measurements instead of groups.)

Measure	Mean	Median	Maximum	Minimum	S.D.
Before	9.6*	8	18	3	4.6
After	13.6*	12	27	7	5.2

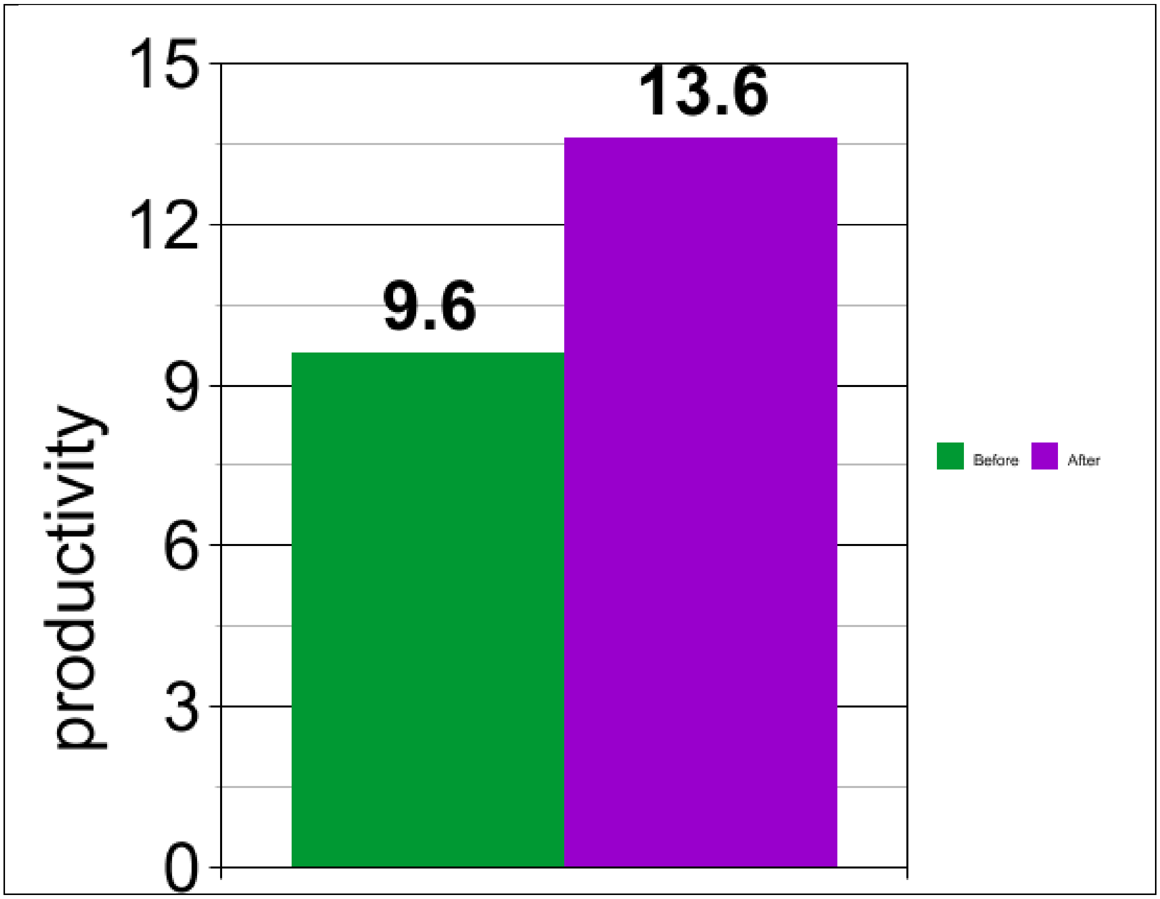
p < .001

An even simpler way to report the results would be to say how many subjects (or what percent) improved.

Notice some differences compared to the separate groups table. There is no need to indicate the sample size of each measure, since they should be the same (the entire sample size, N). Notice that (if the sample size is the same as what we use for a separate groups) significance tends to be better. This is because we are now less worried about inter-subject variance. Reporting confidence intervals or effect size has not yet caught on with repeated measures designs.

The best way to graphically summarize the data in a chart would be to use vertical columns on a bar chart representing percents, medians or means. Each column would represent a different measure, as we see in the following examples.

This bar chart just shows the means of the before and after measures depicted in the above table for interval scales.



You could also have such a comparison chart for the two measures to indicate how they differ as to the median, dispersion (standard deviation or range), maximum, minimum, or percent achieving a certain score or above.

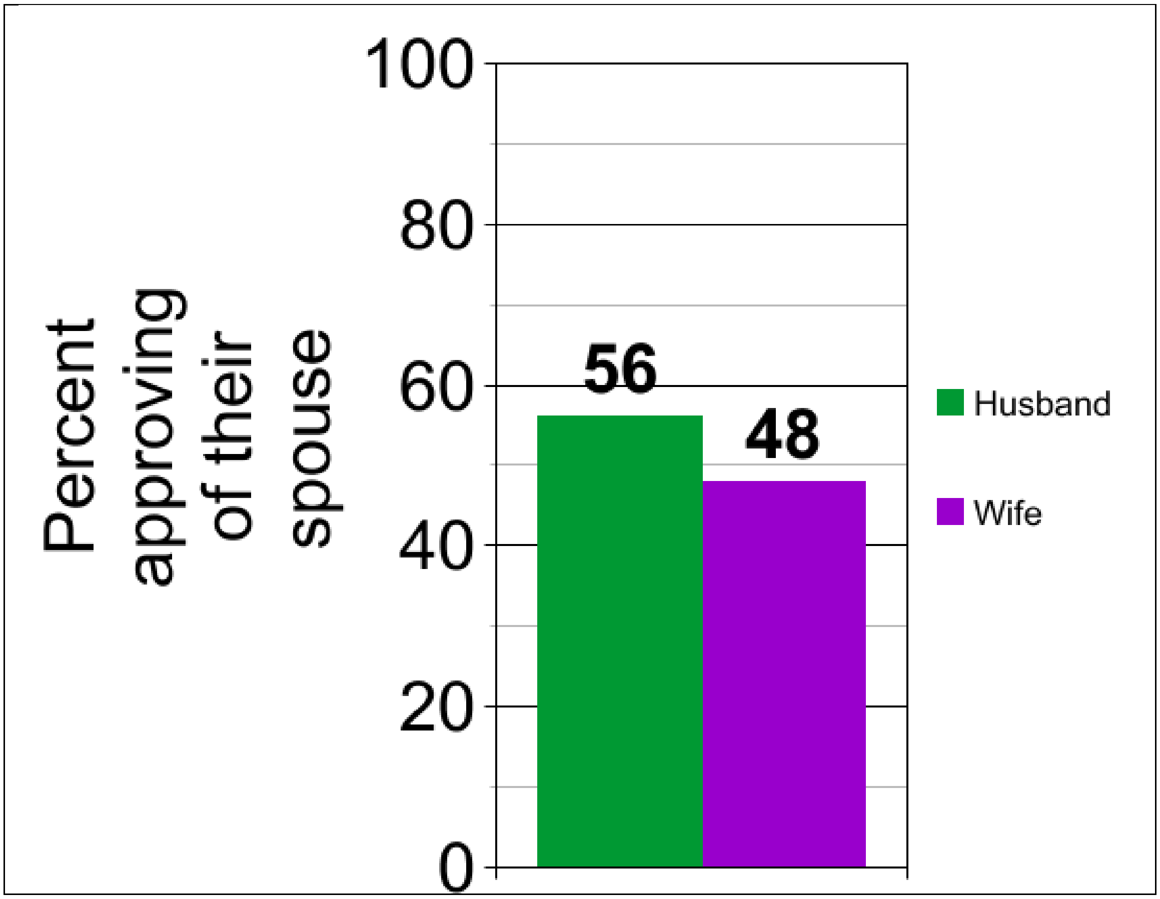
For a criterion variable measured dichotomously, such as marital satisfaction within 50 couples, I like the previously displayed two-by-two contingency table, because it shows how many couples agreed (A = that the spouse is good, or D = that the spouse is not good) and it also shows how many couples disagreed and how many were in each type of disagreement (B = husband more satisfied than the wife, or C = wife more satisfied than the husband).

The following summary table would also work.

Report from	Number approving	Percent approving
Husband	28	56%
Wife	24	48%

p > .10

The bar chart also works to compare the total scores or percents of these measures.



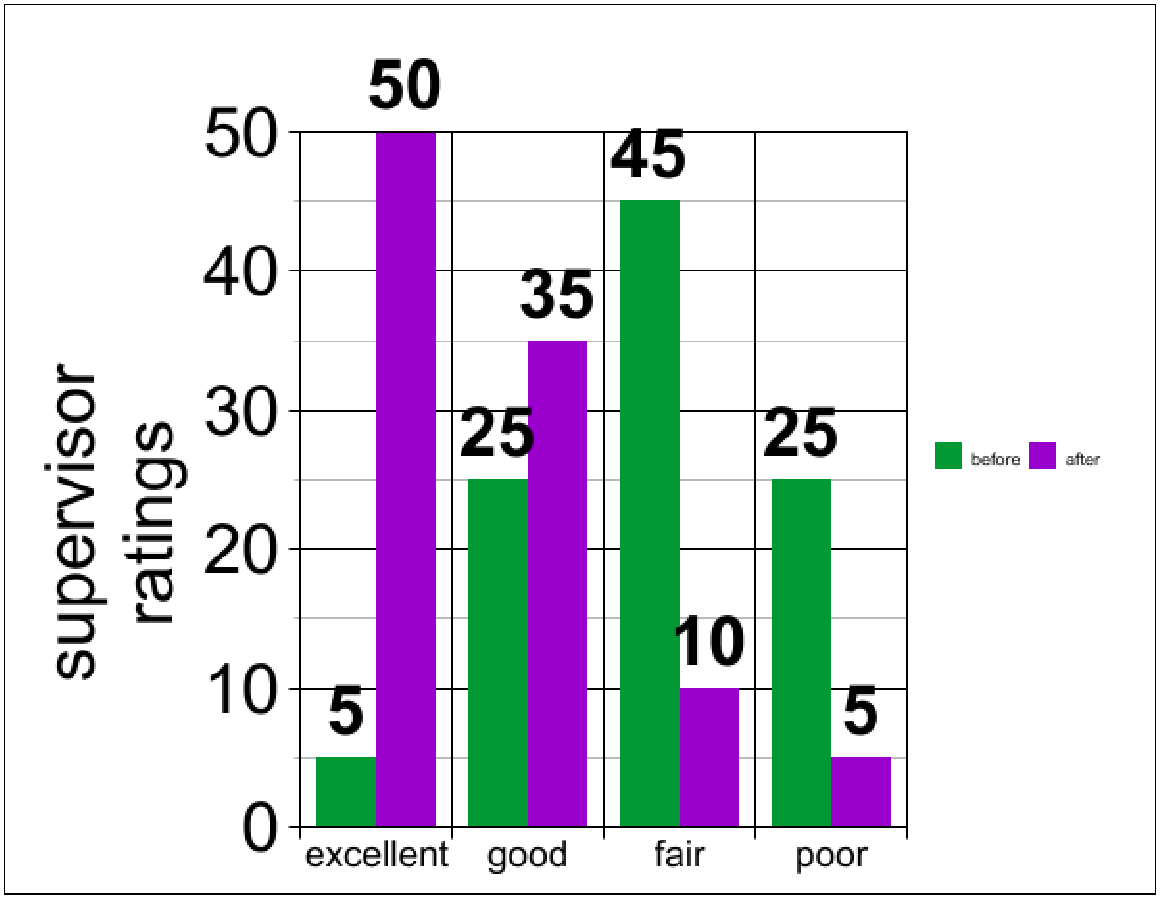
Now suppose that the outcome variable is measured on a multiple nominal scale (or an ordinal scale with four levels). Let’s say that the subjects are salespersons (N = 20). We look how their managers rate them on an ordinal scale (excellent, good, fair, poor) before and after training. The sign test (a special version of the binomial distribution) or the Wilcoxon test could be used as the inferential test. (If you have three or more measures, use the Friedman test.)

Measure	Excellent	Good	Fair	Poor
Before	5%	25%	45%	25%
After	50%	35%	10%	5%

p < .001

An even simpler way to report the results would be to say how many subjects (or what percent) improved. This chart suggests that most did improve, but it could have been as high as 90%.

Again, the bar chart would be the best graphic to employ. Now, it would be better to use four bars for each measure, showing the distribution over that entire variable.



Now, let’s compare all the (single predictor variable) designs we have learned so far, from worst to best. Let our examples come from the field of industrial psychology. The topic will be: “Does special training help certified nurse assistants (CNAs) avoid injuries that they receive from combative patients (e.g., hitting, kicking, biting, spitting)”?

Design	Example
Hypothetical introspection	I was just imagining that if I could be trained to look for non-verbal precursors of patient aggression, I could intervene quicker, or at least get out of the way. I bet it would work.
Actual introspection	I had the training and now I feel more confident, and am avoiding patients who get a little too feisty, and haven’t been hit since. I think the training might be helping.
Case study	I observed one of my colleagues go through the training. She had been hit several times last year, but has avoided a recurrence over the last two months. When I interviewed her, she thought the training might be helping.
Case control or correlational survey	We asked all CNAs who work here (n = 25) if they had gone through the training and if they had been hurt by a patient in the last month. Those who did not go through the training had twice the risk of being hit by a patient.
Experiment after only	We put all CNAs through the training last week, and since then no one has been hurt.
Sample vs. norms	We put all CNAs through the training last week, and since then no one has been hurt. The average for the industry would be two incidents in that time period.
Quasi experiment	We provided the training, and 10 CNAs decided to take it (with the other 15 becoming our comparison group). The trained group had no incidents this week, but the control group reported several incidents.
Within-subjects (before & after)	We put all CNAs through the training last week, and since then no one has been hurt. In the previous week, three CNAs reported an incident.
Randomized Control Trial Experiment	Ten subjects were selected by a lottery and required to attend the training. The other fifteen were not permitted to train at this time. During this week, the experimental group had no incidents, while the control group reported two incidents.

Chapter #11: Multivariate Quantitative Designs

The simplest quantitative designs look at two variables: a *dependent* (effect) and an *independent* (the presumed cause of the dependent). In psychology, the dependent variable is always some *response* that the organism (i.e., subject, participant, patient) makes and the independent variable is some background factor (e.g., age, gender, ethnicity, early childhood experiences) or some current environmental *stimulus* that may have an influence on that dependent variable. A further explanation of these variables can be found in this video.

In a separate groups design, the dependent variable is measured, while it is usually the independent variable that defines the grouping. In a true (*randomized control trial*, RCT) experiment, the grouping is a result of random assignment, and is following by treating those two groups differently (and that different treatment is the independent variable being manipulated). In a repeated measures design, the dependent variable is the one measured twice in an experiment, once before the intentional change (treatment) and then again afterward. In a longitudinal survey, the independent variable would be age and/or the events that take place over the lifespan. In a sample vs. norms design, the norms are for the dependent variable. The independent variable is the one that is supposed to define the difference between the sample and the population, and could be a background factor, stimulus, or even an experimental manipulation.

Complex Relationships

Simple designs are content with measuring (or manipulating) one independent variable, and measuring one dependent variable (as long as other independent variables can be held constant or handled by the randomization of assignment to the groups). This chapter examines more complex designs for additional insights coming from looking at additional variables.

Measuring several outcomes (dependent variables) not only increases the chance that we will find a significant relationship between a given predictor and some outcome, but it will allow us to identify just what kind of impacts (plural) a given treatment or background factor has.

These different outcome measures can be in the form of several related (*inter-correlated*) dependent variables. What is known as *path analysis* may suggest that independent variable A influences *intermediate* variable B which then has an impact on outcome variable C, rather than A having an unmediated impact on C. In such a situation, variable B has a *moderating* role which could serve to *potentiate* (strengthen) the relationship between A and C or *attenuate* (weaken) that relationship.

Multi-item Measures of a Variable

Sometimes the use of several outcome measures is an attempt to provide a more comprehensive operational definition of a single dependent variable, and also one that is more valid and more reliable. Many scholarly journals are reluctant to publish research in which the dependent variable is measured only by the response to a single item, especially one that was developed by the student writing the article.

For example, suppose you have discovered a new and effective treatment for depression. You have a sample randomly selected from a clinical population, and random assignment to a double-blind placebo design, but your only outcome measure is one question you invented "Are you depressed today?" with a binary nominal response: yes or no. That dependent variable measure will be the Achilles heel of your research and would make it most unlikely to be publishable.

If you are only going to use a one-item outcome measure, it would be better to use something more established, such as some of the ordinal response formats used by Gallup and other polling organizations, especially formats with five, seven, or ten levels.

Better yet would be to use an established multi-item test to measure the variable. In the case of depression, this could be one of the many tests that previous clinical studies have used in articles published in the same psychiatric and clinical psychology journals in which you would consider publishing (e.g., Hamilton Rating Scale, Beck Depression Inventory, Geriatric Depression Scale).

Reliability & Factor Structure

The more items that a scale is based on, the more *precise* can be the score of each subject. A patient given the full Geriatric Depression Scale can be classified as "normal range" or "mildly depressed" or "moderately/severely depressed" for clinical purposes, but for research purposes, precise scoring can be from 0 to 30 depressive answers.

Having more items also leads to greater *reliability*: both test/retest as well as inter-rater. Even if there is some inconsistency on one or two items, with more items total, it is more likely that similar scores (consistency) will be the result. Correlation coefficients are used to calculate such reliabilities.

However, multi-item scales open up the question of a different kind of reliability, *internal*: do a subject's answers given on the different items of the scale form a consistent pattern? We could use a Pearson coefficient to calculate how subjects do on one half of the test as well as on the other half. More specialized coefficients (e.g., *Cronbach*, *Kuder*) have been developed especially for internal reliability. A high value indicates that all of the items on the scale consistently measure the same variable. When all the items have a high inter-correlation, and we don't see just a few inter-correlating over her and a different set inter-correlating over there, we have a *uni-factorial* scale. An example of such a uni-factorial depression scale would be the aforementioned Geriatric Depression Scale.

On the other hand, some scales are intentionally *multi-factorial*, and permit the use of several subscales (e.g., the Center for Epidemiological Studies Depression Scale or CES-D measures separate dimensions of depression). Any multi-item scale permits the use of *item analysis*, in which the independent variable can be correlated with each individual item on the test, as well as the score for the aggregate scale and the subscales.

Multi-factorial Research Designs

The strategy in simple quantitative research is to deal with only one independent variable (through manipulation in an experiment or measurement in a survey) while potentially confounding independent variables are either held constant (controlled) or randomized through selection or assignment.

Multi-factorial is also a name for a design in which there are several independent variables. These designs attempt to simultaneously manipulate (or measure) more than one independent variable at the same time. Suppose we measured gender (male or female) and manipulated treatment (experimental or control) we would have a two-by-two (abbreviated 2 x 2) *factorial* design with four resulting groups:

- males in the experimental group
- males in the control group
- females in the experimental group
- females in the control group

Most multi-factorial designs involve separate groups, but some of the factors can be addressed as repeated measures using a within-subjects approach. The between-subjects approach is usually better, but requires a larger sample size. In multi-factorial designs, the groups do not have to be equal in size, but no group should be really small or empty, especially because these designs tend to use parametric statistics, such as the two-way Analysis of Variance (*ANOVA*) or Structural Equations Modeling (SEM).

We could put the subjects in a 2 x 2 table, like this.

	Males	**Females**	**Totals**
Experimental	Men in experimental group	Women in experimental group	**All subjects in experimental group**
Control	Men in control group	Women in control group	**All subjects in control group**
Totals	**All men**	**All women**	**N = total sample size**

This may look like the two-by-two contingency table used for the Yates Chi Square and Fisher Exact Probability Test. However, that approach is used when we are trying to correlate an independent (or predictor) variable we use to define our rows with a dependent variable defining our columns. In the above multi-factorial design, we see the interaction of two independent variables, resulting in the definition of four distinct groups.

We are not limited to 2 x 2. We could add yet a third dimension (variable), such as personality (introvert vs. extrovert) and have a 2 x 2 x 2 design with eight resulting groups. Nor does each grouping have to be binary: we could introduce age as a variable having three levels (under 20, 20-29, and 30+) yielding a 2 x 2 x 2 x 3 design (which would mean comparing 24 different groups). We are only limited by sample size (and the ability to get enough subjects for each group).

One benefit of these multi-factorial designs is that we can look at the impact of each independent variable on the dependent variable(s). But the greatest insight is offered by the capacity to identify the *interaction* between variables. Suppose in the above 2 x 2 factorial design, we are talking about an example from industrial psychology, and the experimental group gets a new form of intense training, while the control group is just supposed to study on their own. If the men respond very well to the training, but the women are turned off by it, we could see the following results: not much difference between experimental and control groups, not much difference between (all) men and (all) women, but a great difference between the four groups indicating an interaction: men with training scored very high while women with training scored very low.

The first time you try a 2 x 2 multi-factorial experiment grouping, just remember that you must end up with four distinct groups. A common mistake by novices would be to put all the males in the experimental group and all the females in the control group. This would mean you would still end up with just two groups, and this would lead to confounding. If you were to find a difference between those two groups, you would not know if that were due to the variable of gender or the variable of the training, as demonstrated in this video.

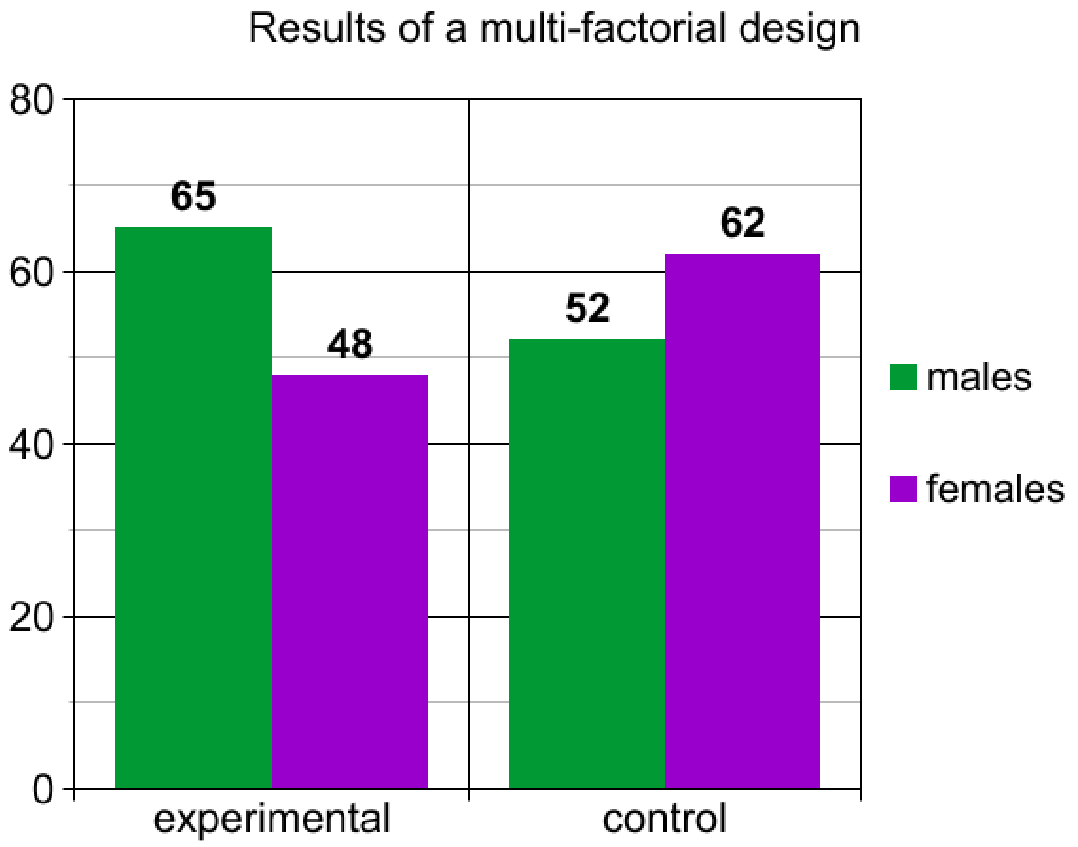
Multi-factorial Statistics

The parametric statistical test used for comparing these groups would be ANOVA (Analysis of Variance). A *one-way* ANOVA could tell us the significance of the difference between the means of these groups. A *two-way* ANOVA also looks at the interaction between the independent variables. A *MANOVA* is appropriate when we have multiple dependent variables as well. *ANCOVA* and *MANCOVA* would be used when we also want to look at the *covariance* of these. All of these tests are parametric and make assumptions such as normality of dependent variable distributions, independence of observations (e.g., random assignment), and homogeneity of variances and covariances and *sphericity*. If these assumptions are violated, *Type I* errors become more likely.

When all of these parametric assumptions cannot be met, the solution is to use a nonparametric alternative of ranks based inferential tests. For a separate groups design, the *Kruskal-Wallis* test can be used. For a repeated measures design, a *Friedman* test can be used. Both of these are based upon chi squared (and thereby have those limitations).

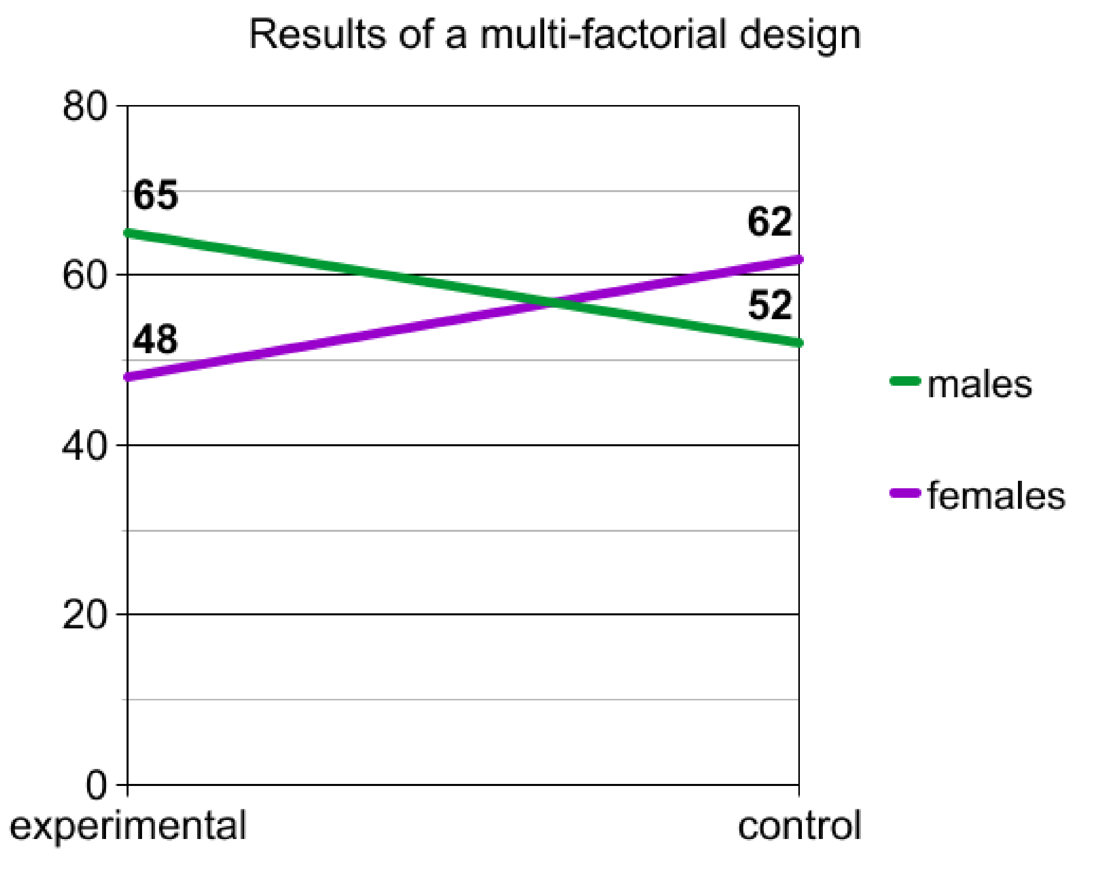
Multi-factorial designs use often use *bar* graphs to display the results. The number of bars is the number of groups, while the height of the bars is the measure of the dependent variable's central tendency (e.g., percent, mean, median). Don't use absolute numbers of subjects in each group to determine the height of the bar, because the group sizes may be different.

Use of Bar Chart in Multi-Factorial Design



More frequently utilized are *line graphs*, where the difference between the *slopes* of these lines represents the interaction of the independent variables. This is most visually stunning when the two lines crisscross, but that is not necessary for there to be an interaction depicted. All we need to have is different slopes for the two lines. The greater the difference between the slopes, the greater the interaction between the independent variables.

Use of Line Graph in Multi-Factorial Design



Correlational Designs for Multiple Variables

A correlational design can also have multiple independent variables and/or multiple dependent variables. We have already seen how a *correlation matrix* can represent all the correlations between all possible combinations of variables.

Here is what a correlation matrix for five variables might look like. Gender, age, and grade in school are clearly predictor variables. IQ can be viewed as another predictor variable. Scores on academic tests would be the criterion variable.

	Gender	Age	Grade	IQ	Scores
Gender	1.00	.06	-.01	-.12	.04
Age	.06	1.00	.87 ***	.10	.38 *
Grade	-.01	.87 ***	1.00	.13	.21
IQ	-.12	.10	.13	1.00	.51 **
Scores	.04	.38 *	.21	.51 **	1.00

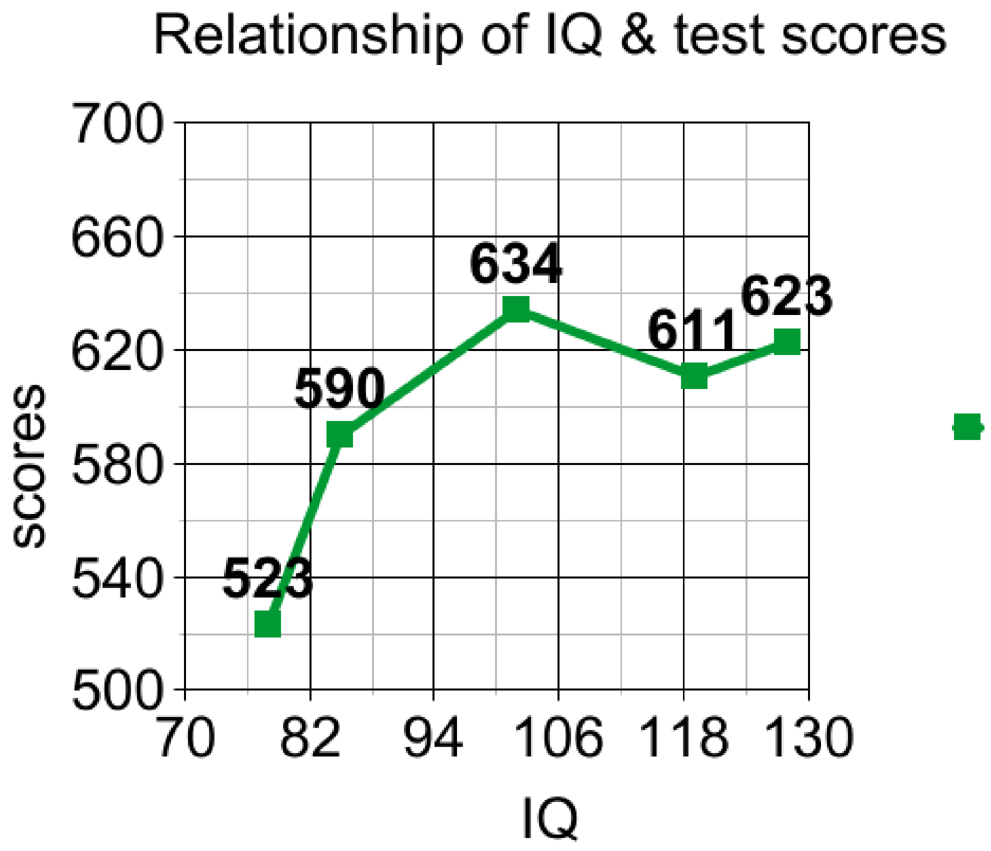
* $p < .05$ ** $p < .01$ *** $p < .001$

When we are just correlating two variables, one predictor and one criterion, we can create a *scatterplot* showing the distribution of these variables, based upon an equation of

$$Y = AX + B$$

where *X* is the predictor variable and *Y* is the criterion. The letter A represents the *slope* of the line and the letter B represents the *intercept* of the line on the Y axis.

Use of a Scatterplot for Predictor and Criterion Variables

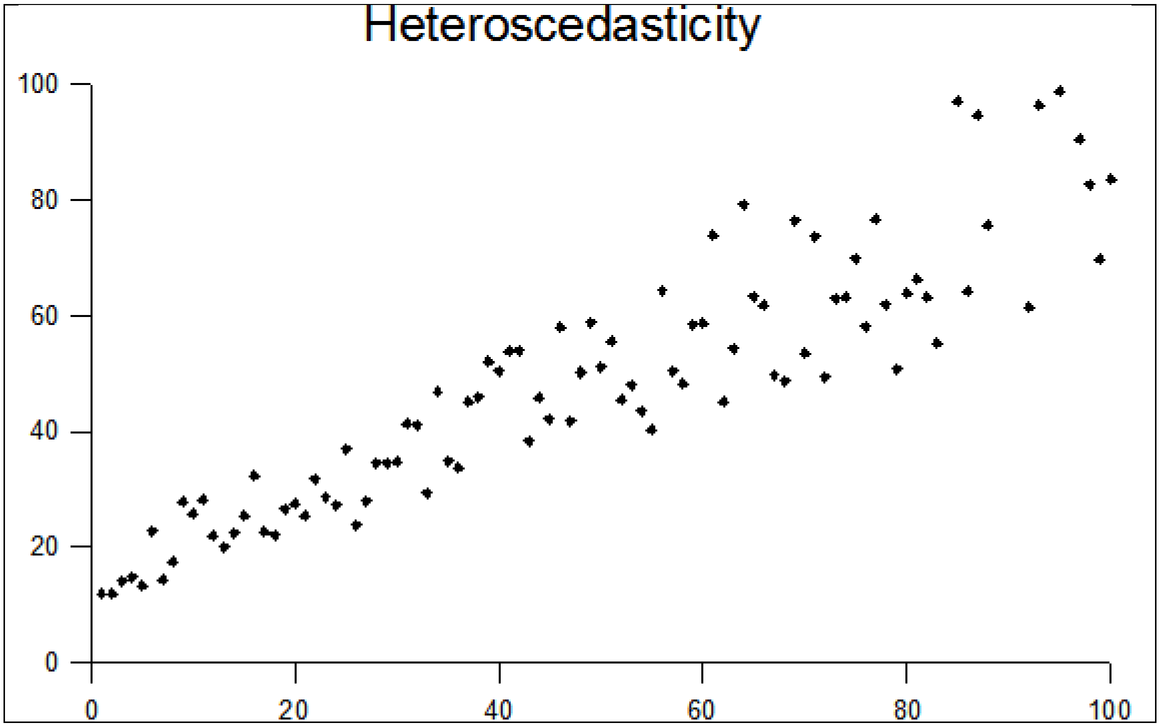


When we have several independent variables we, are using multiple regression and we have to consider the possibility of *multicollinearity* (in which some of the independent variables are correlated in a linear fashion). This means that one predictor variable may be the real predictor of (perhaps by having an impact on) the dependent variable, while the other predictor variable has no real impact on the dependent variable (but nevertheless can still predict the dependent variable). Another possibility is that each predictor variable contributes something to the prediction of the dependent variable, but multicollinearity means that this is not simply additive. A *multiple regression* equation of two predictor variables X and W might look like this:

$$Y = AX + CW + B$$

The above equation assumes that both predictor variables have *linear* relationships with the criterion variable. If the relationship is non-linear, the equation may have to use powers (e.g., squares, cubes) or logarithmic transformations of the predictor variables. The Statcato program is especially user-friendly for developing different kinds of multivariate equations.

Another concern with multivariate designs is *heteroscedasticity* which means that the strength of the correlation varies over the range of a variable. The opposite is *homoscedasticity*, which is consistency of a correlation across a range. For example, look at the correlation between age and crystallized intelligence. The correlation is positive and strong for the first two decades of life: each year brings more knowledge so that the average 16-year-old knows much more than the average 10-year-old. But that correlation evaporates after about age 20. It is not so much that the curve just flattens out, but that the variables are just not correlated: some people don’t learn much after age twenty, while many continue to learn new things, so the scatterplot would show a linear relationship only over a segment of the range.



Final Recommendations

Unless you are using innovative techniques with a hot new topic, your research will have to embrace multivariate techniques in order to be publishable. However, don't bite off more than you can chew. To do this well requires larger samples than most students can conveniently access.

Chapter #12: Other Samplings & Codings

Rows & Columns

Our previous units assumed that we could put the data into a spreadsheet such that each subject would occupy one row and each variable would occupy one column: one subject per row and one row per subject; one variable per column and one column per variable. The number in a given cell (e.g., 11C) would represent the numerical score (perhaps dummy coded) of subject #11 on variable C. The ideal situation is for all of that subject's data to be only in row #11, and all of the data on variable C to be only in column C. (The only real exception to this was in the repeated measures design known as *matched pairs,* and in that situation we still used these rows and columns guideline, except that a row contained data from the pair of related individuals (e.g., husband/wife, matched controls).

The above coding situation of rows and columns in a spreadsheet usually can be accomplished easily when we enter the data, subject by subject, from the raw data of questionnaires because all the data for that individual subject should be right there on the questionnaire: background factors, attitudes and performance. If there is some other important variable that needs to be included in the survey, but is not going to be measured by the subject's responses on the questionnaire, then the examiner needs to code that information right on the questionnaire.

For example, my very first survey was done fifty years ago when I was a freshman at Claremont Men's College and took my first psychology class. Although CMC was an exclusively male college (then), it was in a consortium with other co-ed colleges (e.g., Pomona, Pitzer) and a women's college

(Scripps). A main topic of conversation in the dorms were the (alleged and fantasized) differences between Pitzies and Scrippsies. So, that became the topic of my first survey. An art major sophomore and I developed a series of about ten questions describing attitudes and personality traits on which we had hypothesized a P-S difference. The best part of this project was walking around the women's dorms recruiting subjects. Here's how this illustrates my point. The questionnaire did not have two important variables printed on it: the subject's college and dormitory. We had to write that on the back of each questionnaire so we did not forget. (Some of the girls may have worried that we were writing down notes like "Cute, call her later" but we never followed up anyone for a date.)

Easy rows and columns coding is also derived from most *archival* data (e.g., patient charts, student files, job applications). Indeed, some of these records are already in a spreadsheet or other database format and you just have to do a simple download or copy and paste. Of course, with such organizational records, you must make sure that your access and use of data

- follows organizational policies
- has the approval of someone in authority
- protects the anonymity (or confidentiality) of the subjects
- does not violate HIPAA or FERPA laws

However, this spreadsheet model for coding is not always necessary, and sometimes it is not even possible. These situations (constraints or opportunities, depending upon how you look at them) are most likely to occur when we are not using questionnaires or archives as our data sources.

Just Record One Variable? (binary nominal scaling)

Suppose your only source of data is a *field count*: a simple (usually public) observation of the subjects. The population might be workers in their place of employment, commuters on a subway train, drivers at an intersection, pedestrians crossing the street, students in a classroom, customers waiting in line for a store to open, fans in a stadium, kids playing AYSO soccer on a Saturday morning, worshippers coming out of a church, or people in a park.

What distinguishes a field count from other kinds of naturalistic observation is the emphasis on quantification. Ethnographies and participant observations also look at people in their normal environments, but those qualitative studies yield narrative data only.

Suppose we have a field count and the only variable I am able to identify about the subjects is the gender of each one of them. Theories guiding the formulation of my hypotheses would involve reasons for the supposed differences between males and females (e.g., genetics, child-rearing practices). The dependent variable would have to be that the subjects chose to be present at the location at that time. The simplest kind of design would be *sample vs. norms*. Here are some examples of research questions that could be explored relevant to these theories, utilizing field counts with sample vs. norms designs.

- Are psychology courses disproportionately female? Go into the classroom and count the number of males and the number of females. Is it close to evenly divided?

- Are the customers at Redlands Sewing Center disproportionately female? Wait across the street between ten and eleven on a Thursday morning and count how many women and men (not many) enter the store.

- Are visitors to the Christian Science Reading Room (across the street from the Smiley Public Library) disproportionately female? Count how many men and how many women go in. (If you only spend an hour, you may not get enough of a sample size.)

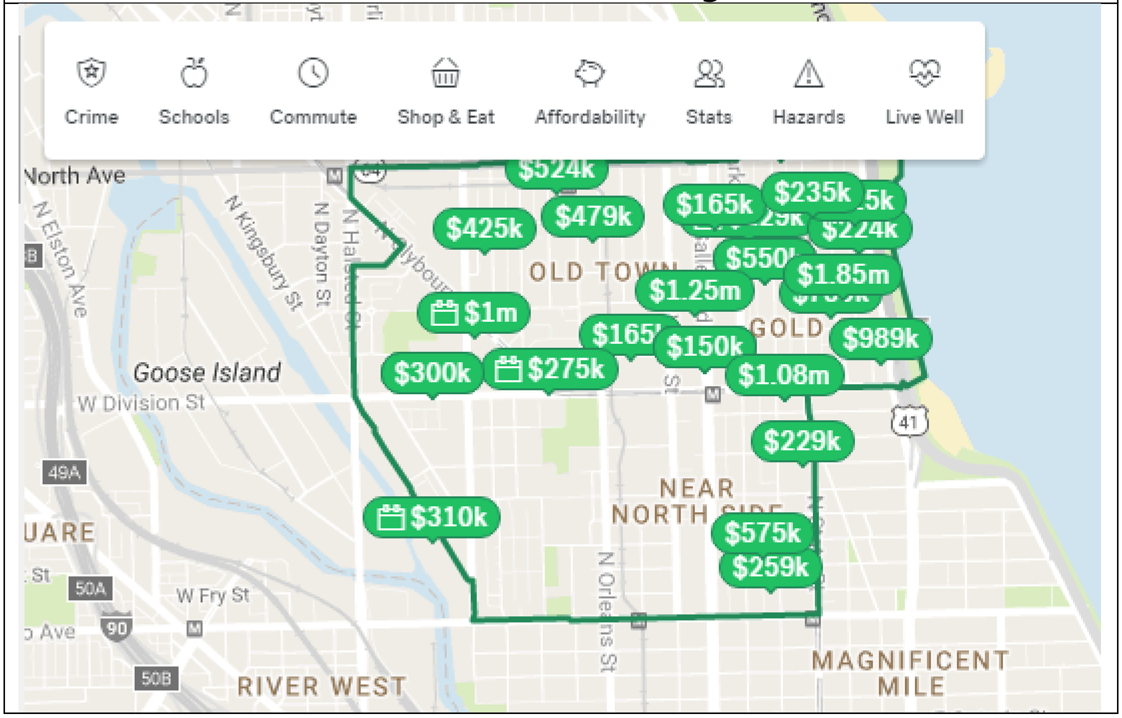
- Are skateboarders disproportionately male? Go to a skateboard park in Long Beach and notice 13 youths skating around: all boys. Here's the video.

In each of the above examples, we have a binary nominal scale (male/female). We could assume that the population norms are half male and half female, and then employ the *binomial* distribution as our inferential statistic.

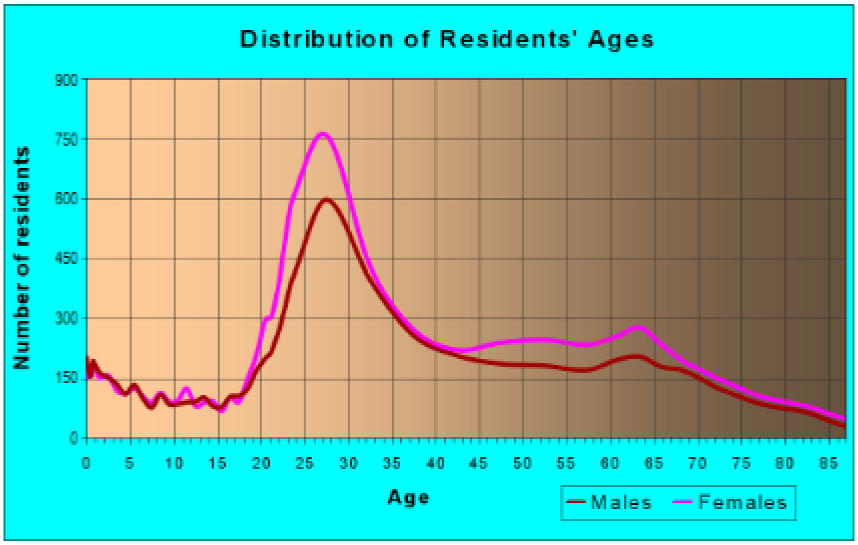
Of course, that 50/50 split is theoretical. There are situations where the population has a more lopsided male/female split. For most community colleges, the norm is closer to 53% female (but ranges from about 45 % - 60%, depending upon the college's location and course offerings. Here are the current ratios at Crafton Hills College.

One of the easiest places to access neighborhood data about such variables as gender would be the real estate site, Trulia, where you select a city (Chicago) or zip code (60610) or neighborhood (Gold Coast). The first thing that comes up are housing prices, but you can search specific information about crime, schools, age, and education levels.

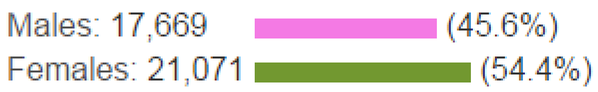
Trulia information about Chicago’s Gold Coast



City-data gives even more demographic information. Gold Coast is a neighborhood of young professionals.

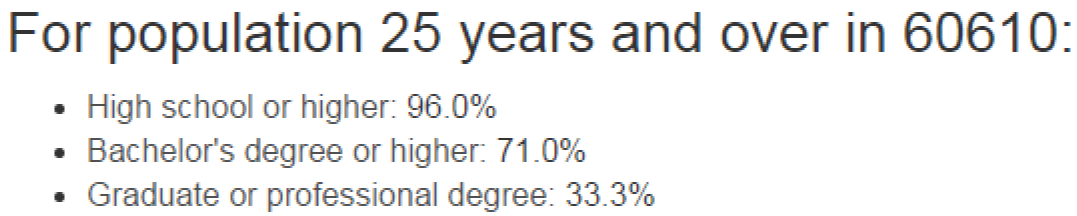


There are slightly more women than men, and the education level is high.



Males: 17,669 (45.6%)

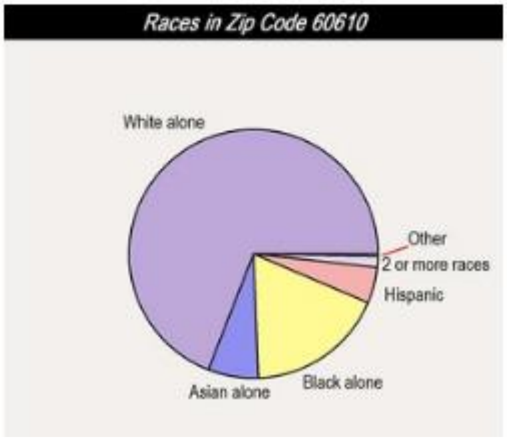
Females: 21,071 (54.4%)



For population 25 years and over in 60610:

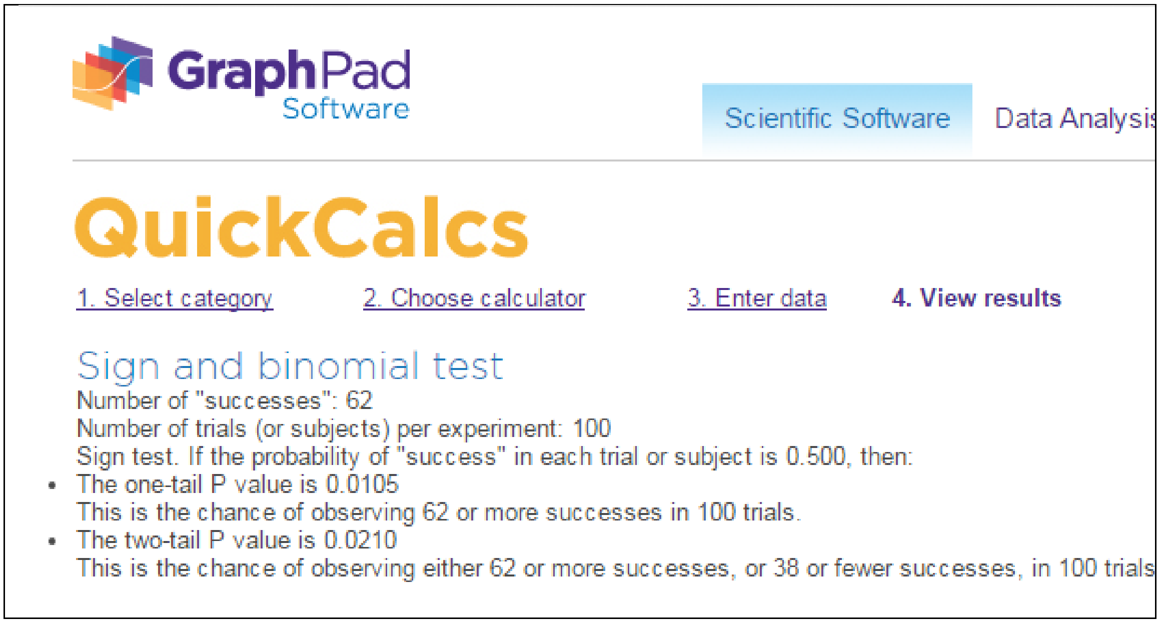
- High school or higher: 96.0%
- Bachelor's degree or higher: 71.0%
- Graduate or professional degree: 33.3%

Compared to other Chicago neighborhoods, it is disproportionately White, under-representing both African-Americans and Hispanics.



One of the challenges of using sample vs. norms designs for surveys (or for experiments, as in chapter #10) was deciding which norms to employ, especially in geographical and organizational studies. For example, my college might be 54% female in terms of overall number of students, but women might be 56% of the day students (when the observation was made). If women are taking more classes, they might be 61% of the bodies in the classes overall.

This approach of binary nominal measurement of gender and a sample vs. norms design can also work with archival data. The topic question might be: Are women underrepresented in the highest ranks of financial advisors. A recent issue of *Barron's* (July 20, 2015) listed the 100 top financial advisors in the U.S. Out of this sample, we could go through the list and categorize each name as likely male or likely female (except for a few like Koo or Jordon): 62 were definitely male names. Now, we turn to the *binomial* distribution as our inferential statistic.



Sign and binomial test

Number of "successes": 62

Number of trials (or subjects) per experiment: 100

Sign test. If the probability of "success" in each trial or subject is 0.500, then:

- The one-tail P value is 0.0105
  This is the chance of observing 62 or more successes in 100 trials.
- The two-tail P value is 0.0210
  This is the chance of observing either 62 or more successes, or 38 or fewer successes, in 100 trials

These data have fair significance (p < .05), so the null can be rejected.

The most essential thing to have with any sample vs. norms design would be the norms (usually from a population). The topic of gender distribution works nicely within these constraints: something observable in a field count, binary nominal scaling, available norms. But gender is not the only topic that can be investigated within these constraints.

Whenever we see the distribution of something in two time periods or two locations, or two random outcomes, we could consider this approach. Here are some examples.

- For this semester, is the on-campus section of Psyc 201 more popular than the online section? Count up how many students registered for each class.

- On Sunday morning, is the Stater Bros. Market on Lugonia & Wabash more popular than the one at Colton & Orange? Count up how many customers go in each market between 9 AM and 10 AM (on the same day of the week).

- Do gamblers on the roulette wheel tend to make more bets on black or red? Watch one table for an hour and count how many gamblers bet red and how many bet black.

- Do the birds swimming in the fountain prefer swimming in the end where the water comes out? The fountain has two sides and we could count how many birds are in each side (as seen in this video).

For each of the above examples, the “norm” would be a *random* distribution of 50% in each of the two possible categories. If there is no trend for students to prefer the online class over the on-ground class, then we would expect classes to have close to equal registration. If there is no trend for one Stater Bros. market location to be more popular than the other, then we would expect both to have close to the same number of customers. If gamblers have no real tendency to prefer red or black, then we would expect to observe close to an equal number of bets. If the birds don’t have a particular preference for one end of the fountain over the other, we would expect close to an equal number of birds at each end.

One of the biggest mistakes students make when conceiving the type of research represented by the above examples is to assume that the design is a separate groups comparison: this class vs. that class, this location vs. that location, red bets vs. black bets, a group of birds on this end vs. a group of birds at that end. In most real separate group designs, the grouping is done on an independent variable, either a manipulated one (i.e., an experiment) or a background factor (e.g., religious denomination). Especially in the above four examples, the presence of the subjects in each category represents the subjects' choice (and therefore would be a dependent variable). The design for testing the hypothesis is to take the sample and compare it to the norm (i.e., 50%).

Field counts can also use *concurrent* measures and the sample vs. norms design of 50%. Is it true that in most (male/female) couples the male is taller? Look at twenty heterosexual couples walking together in a public place. Note how many times the man is taller than the woman. Now go to the Binomial Test and assume a norm of .5. The biggest problem with this field count would be trying to decide who is a "couple" and who is not. The relationship between two people is hard to infer when the only information about them in a field count is a few seconds of observation, not supplemented by any questions or archival data. (I frequently take my 89-year-old mother for a walk, holding her hand in case she falls, and numerous times people have mistaken us for husband & wife: "Such a cute couple, how long have you been married"?)

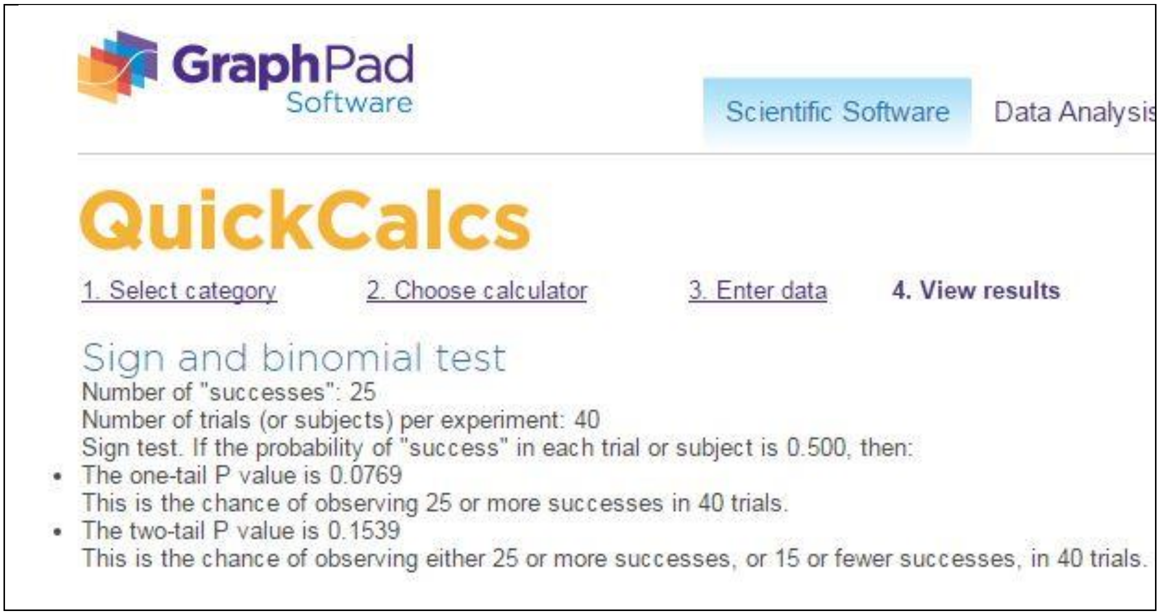
The way to report data from these studies would be to use a simple part / whole percent: we take the part of the sample that is male and divide by the total sample size. The inferential statistic would be the binomial distribution.

For example, if there are 40 students who signed up for a section of the Research Methods class (25 in the day section and 15 in the online section). We would report the statistics like this:

62% registered for the on-campus section (100 X 25/40)

38% registered for the online section (100 X 15/40)

According to the binomial distribution's two tail test, this proportion does not differ significantly from pure chance (so we cannot reject the null).



Sign and binomial test

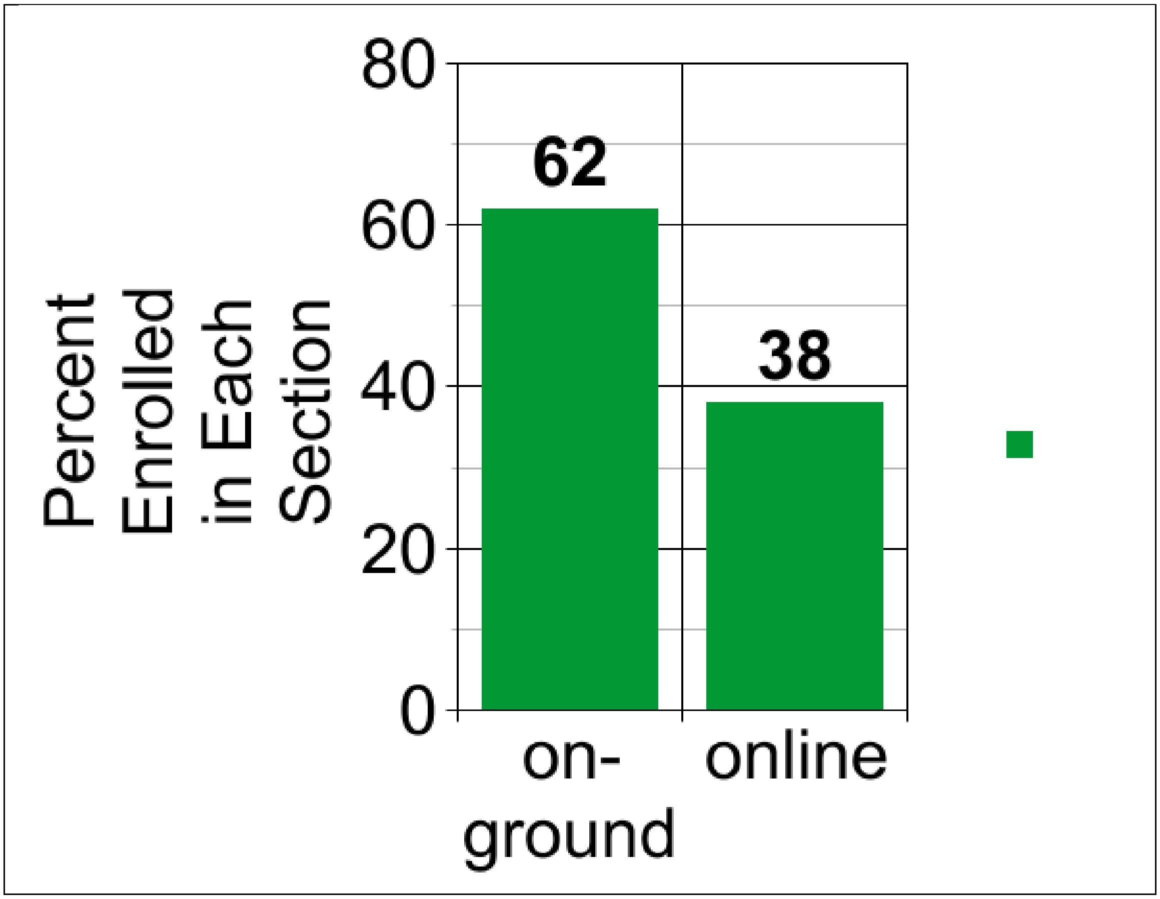
Number of "successes": 25

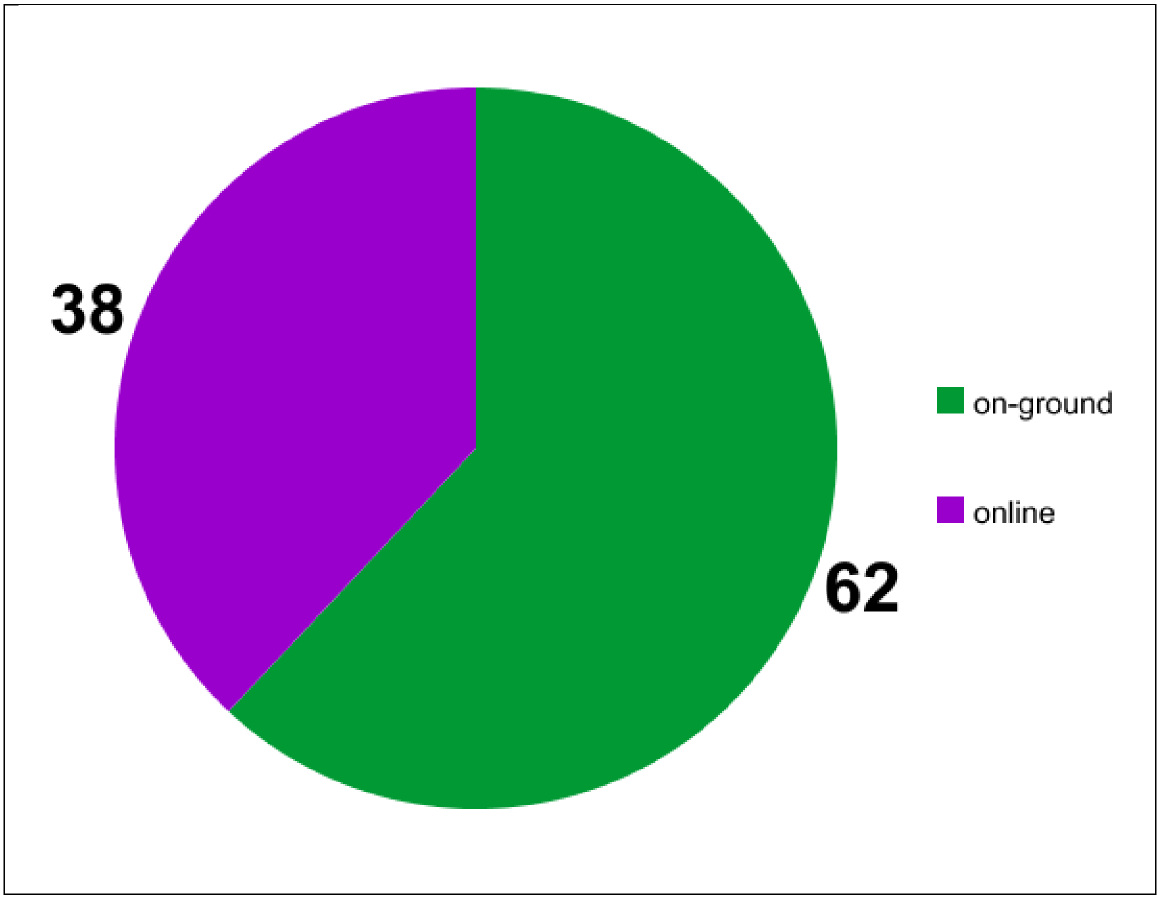
Number of trials (or subjects) per experiment: 40

Sign test. If the probability of "success" in each trial or subject is 0.500, then:

- The one-tail P value is 0.0769
  This is the chance of observing 25 or more successes in 40 trials.
- The two-tail P value is 0.1539
  This is the chance of observing either 25 or more successes, or 15 or fewer successes, in 40 trials.

The results could be visually represented by a *bar chart* or even a *pie chart*, which could be easily done at this site.


Percent Enrolled in Each Section
80
60
40
20
0
62
38
on-ground
online
🟩



Just Record One Variable? (other scaling)

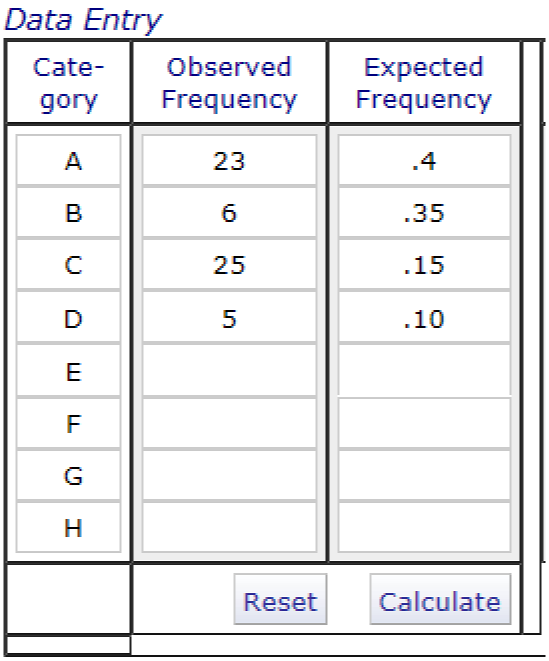
So far we have just looked at norms that were measured in a binary nominal scale: male or female, this location or that location, this choice or that choice. The norms can also be in a multiple nominal or ordinal scale (with just a few levels). Ethnicity is a good example of such a norm.

Let’s do a field count of the patients entering the emergency room at St. Mary Medical Center in downtown Long Beach. We observe each admission in our sample (n = 59) and classify that patient into one of four ethnic categories. To get the percents, divide the part of the sample that fits in a given ethnic category by the sample size.

Ethnicity	Hispanic	White	African-American	Asian
n	23	6	25	5
%	40%	10%	42%	8%
City-wide (2014)	40%	35%	15%	10%

$p < .01$

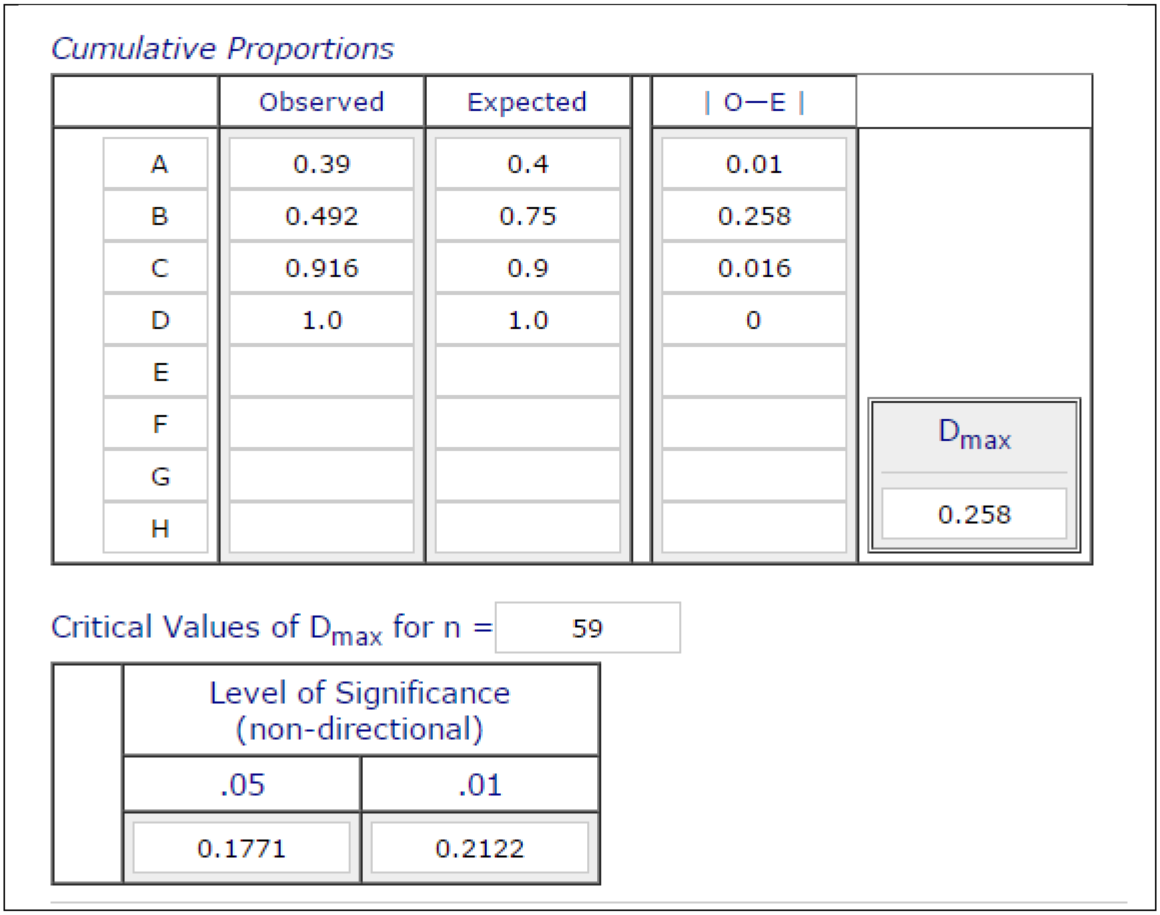
The inferential statistic when there are more than two categories (or levels) would be the one-sample *Kolmogorov-Smirnov* test for the absolute maximum cumulative difference of frequencies. You can find this at the VassarStats site, then click on frequency data and then scroll down to Kolmogorov-Smirnov. Your input would look like this. Notice that the percents in the population are entered as decimal points of expected frequency.



Data Entry

Cate-gory	Observed Frequency	Expected Frequency
A	23	.4
B	6	.35
C	25	.15
D	5	.10
E		
F		
G		
H		

And this is what the output would look like. Notice that both the observed and expected are given as cumulative frequencies. What is the maximum difference at any point between those frequencies?



Cumulative Proportions

	Observed	Expected	| O−E |
A	0.39	0.4	0.01
B	0.492	0.75	0.258
C	0.916	0.9	0.016
D	1.0	1.0	0
E			
F			
G			
H			

D_{max}

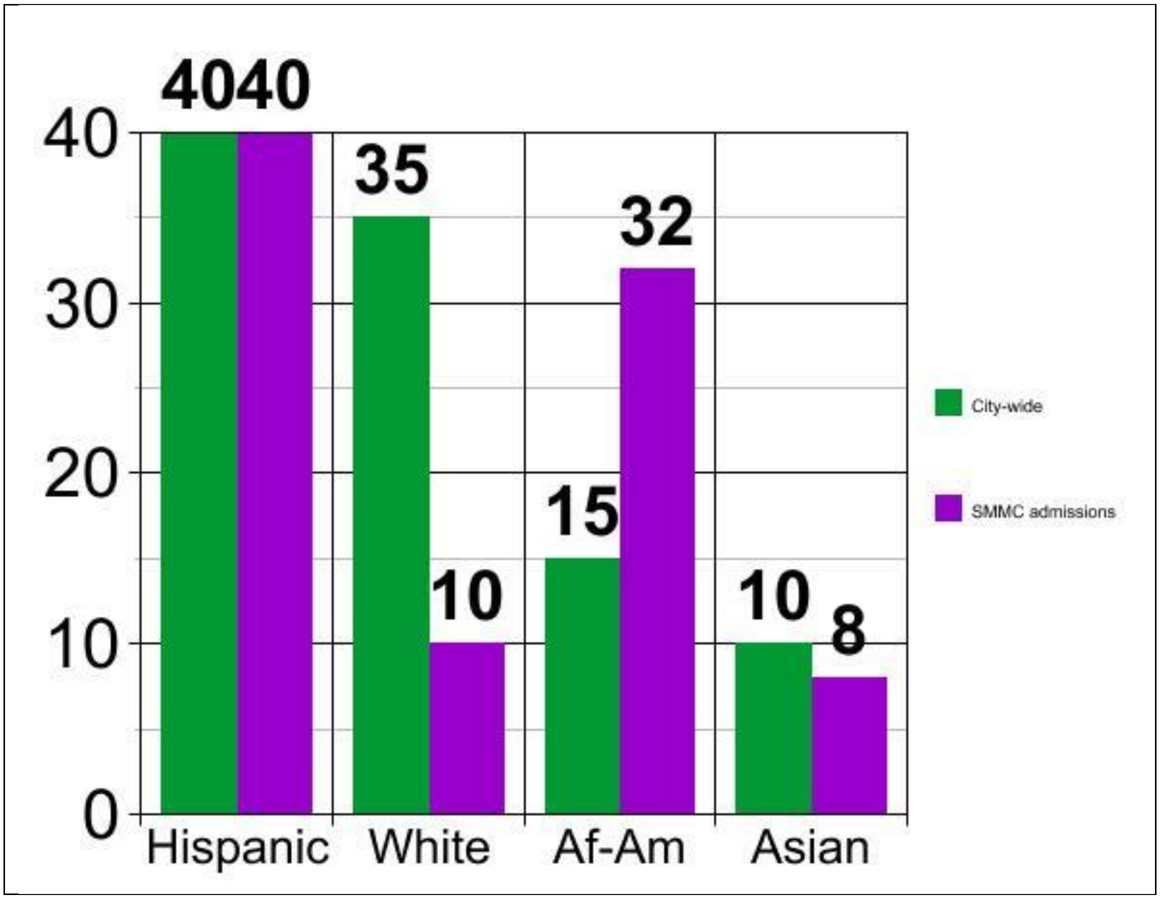
0.258

Critical Values of D_{max} for n = 59

Level of Significance (non-directional)	
.05	.01
0.1771	0.2122

We needed a difference greater than 0.2122 to qualify at the p < .01, level and the difference was even larger at 0.258, so these results differ significantly from the city's norms. While Asians and Hispanics are proportionately represented in the hospital's emergency admissions, Whites are underrepresented while African-Americans are overrepresented. However, this represents a difference between the sample and city-wide norms. If we looked at the population breakdown of downtown, those norms would be closer to the observed sample.

The bar chart is ideal for showing the differences of sample and population across multiple categories.



If there are variables where we could also do a field count using a ratio or interval (probably discrete) scale, the descriptive statistic would be a comparison of means or medians. The means could be compared via a one sample t-test (which Excel could do) but that would entail putting each subject into a row and entering the observed variable in a column.

We could also report the results as the percentage of the sample that was above the median of the population (and use the binomial distribution to see if that sample proportion is significantly different from the population norm of .5, since by definition 50% of the population is above the median).

The great weakness of these sample vs. norms field counts would be the same as the sample vs. norms experiments: the vulnerability to confounding variables. That makes it hard to interpret why there’s a significant difference. For example, if we find that the Stater Bros. Market location on Wabash & Lugonia gets significantly more customers than the one on Colton & Orange, is that due purely to location? There are other important factors, such as the size and age of the store. Even if we can identify location as the major reason for the difference, is that due to the type of customers who live in the respective neighborhoods or is it due to the competition of other stores that are close by?

Type of sampling	**Subjects**	**What to put in rows on the spreadsheet**	**Inferential statistic**
Field count	Organisms who show up to be counted	No spreadsheet needed for one variable measures	Binomial distribution or Kolmogorov-Smirnov
Time period, length, area, space	Organisms who show up to be counted or events that occur	No spreadsheet needed for one variable measures	Poisson distribution
Trace	Organisms not observed but there is evidence of their behaviors	No spreadsheet needed for one variable measures	Binomial distribution or Kolmogorov-Smirnov Or Poisson distribution
Aggregates	Units described by data: schools, cities, companies, states, countries	Units described by data: schools, cities, companies, states, countries	Depends upon scale and design
Small n	Patients	Each subject, but run descriptive statistics by rows	Separate groups (treating all of a subject's "before" as one group and "after" as another

Time Periods & Sample Spaces

Sometimes we don't have an opportunity to select specific subjects to be in the sample (and sometimes we don't even know how many subjects were in the sample). The sample space might be a time period, or a unit of distance, area, or volume. Here are some research questions fitting this pattern.

- Is the number of a bank's customer arrivals significantly higher during the last hour?

- Are there more accidents on the stretch of road (16 miles) between Gilroy and San Jose?

- Are there more gopher holes in the vegetable garden (just one acre) than in other parts of the farm?

- Is more cotton dust per cubic meter found in the new building at the textile mill?

To calculate a p value we need to have the norm for the mean or expected frequency:

- the average (mean) number of bank customers expected in an hour

- the average (mean) number of accidents expected on a 16 mile stretch of highway in California

- the average (mean) number of gopher holes in an acre of that farm

- the average (mean) number of specks of cotton dust in a cubic meter of space in that entire textile mill

Then plug that figure into the *Poisson* distribution equation, which can be calculated at this site. This will tell us what is the probability of getting exactly our result (by pure chance) and the probability of getting that result or fewer.

Traces

A related approach is *trace* research. Here is a situation where we may not even see the subjects, or know who they were, or how many they were. All we have is some trace that the subjects (or their activity) were present or did something.

Trace observed	Organisms (not observed)	Behavior producing trace
Tracks in snow	Deer	Walked by last night
Bird droppings	Birds	Had been on statue
Worn carpet	Museum patrons	Had walked in front of exhibit
Empty supermarket shelf	Shoppers	Had purchased all the items on that shelf
Empty toilet paper roll in last stall of the bathroom	Users of the bathroom	They choose the last stall more often than the others
Beer bottles	Party goers	They choose to drink those brands of beer
Used condoms in the garbage	Residents of the apartment building	Having safe sex
Stack of newspapers for recycling	Residents of the apartment building	Subscribing to newspapers
Parked cars in the shade Pathways across the grass	People who came to the shopping mall today Students	Not wanting to leave the car in the hot sun Students getting to class walked on grass

This video shows trace analysis of chile preference.

Most of these above examples would use a sample vs. norms design, and make null hypothesis statements like each statue will be equally appealing to the birds, each party will consume a similar amount of beer, each apartment building will use an equal amount of condoms and beer and newspapers. The confounding variables are numerous, as it is hard to figure out why observed traces differ. One of the great weaknesses of the trace design is we usually don't know how the numbers of subjects might be distributed between the different sites, because we don't observe the subjects, just the traces of their presence and behaviors.

We only observe that one party resulted in more empty beer bottles than did another party at a different location (or time). Perhaps one of the parties had more guests. Other confounding variables could be the time of year, age of the guests, presence of other liquor, or purpose of the party. For example, this party was in Acapulco and resulted in many empty beer bottles. This wedding celebration in Toluca had much fewer beer bottles the next day. The DJ claimed it was due to the fact that he gave everyone a mask and a balloon, and this served to lower inhibitions and get people to dance without having to drink so much. Consider the possible confounding variables. Did Acapulco have more people attend the party? Was the temperature difference a factor (Acapulco is always warm, Toluca is always cold). Could it be that the wedding had more tequila competing with the beer?

If we cannot observe the individual subjects, we have to stick with the sample vs. norms design. We count up the total number of beer bottles from both parties, find that 58% of the bottles were consumed in Acapulco, and compare that to the null hypothesis of a 50% split.

Another problem with traces and archival data is that we may not be able to observe all of what has been left behind (or even a large portion of it) and probably not a representative sample of it. For example, there is a project to record gravesites around the world. Despite the name of the project, most gravesites have not been registered. I know that my grandfather, George Brink, died around 1957 in New York State, but I cannot find his records. I wouldn’t expect to find my father’s grave on that site, even though he passed away in 2011, because he was cremated.

Separate Groups Field Counts

A field count can also use a *separate groups* design, but you will have to record two variables for each subject: the grouping variable (independent or predictor) and the outcome variable. For example, observe the four-way stop at Church Street and Brockton Avenue in north Redlands. Let’s take the driver’s gender as the independent variable and whether or not the driver decides to make a complete stop at the dependent variable.

Contingency Table for Field Count			
	DV = full stop	*DV = incomplete stop*	*Totals*
IV = Male driver	A	B	A + B
IV = Female driver	C	D	C + D
Totals	A + C	B + D	N = A+B+C+D

A + B + C + D will represent our N, the sample size.

The marginal A + B represents all male drivers coming through that intersection.

The marginal C + D represents all female drivers coming through that intersection

The marginal A + C represents all drivers who came to a full stop

The marginal B + D represents all drivers who did not come to a full stop

Cell A represents the male drivers who stopped

Cell B represents the male drivers who did not stop

Cell C represents the female drivers who stopped

Cell D represents the female drivers who did not stop

You can also use a measure of trace results, without watching the subjects actually perform the behavior, in a separate groups design. For example, you could intentionally drop a stamped, addressed envelope around several dozen different mailboxes around the city. Half of the letters might be addressed to one organization (e.g., a Methodist church) while the other half might be addressed to another organization (e.g., a mosque) to see if the (unobserved) people who find the letters are more likely to help one group of letters get to their intended destination. These data could be analyzed with the two-by-two contingency table

Contingency Table for Trace Field Count			
	DV = letter delivered	*DV = letter not delivered*	*Totals*
IV = addressed to Methodist Church	A	B	A + B
IV = addressed to a Mosque	C	D	C + D
Totals	A + C	B + D	N = A+B+C+D

A + B + C + D will represent our N, the sample size.

The marginal A + B represents all the letters distributed that had been addressed to the Methodist Church

The marginal C + D represents all the letters distributed that had been addressed to the Mosque

The marginal A + C represents all those letters that had been delivered, regardless of the address that appeared on them

The marginal B + D represents all those letters that had not been delivered

Cell A represents the letters delivered to the Methodist Church

Cell B represents the letters that had been addressed to the Methodist Church, but did not arrive.

Cell C represents the letters delivered to the Mosque

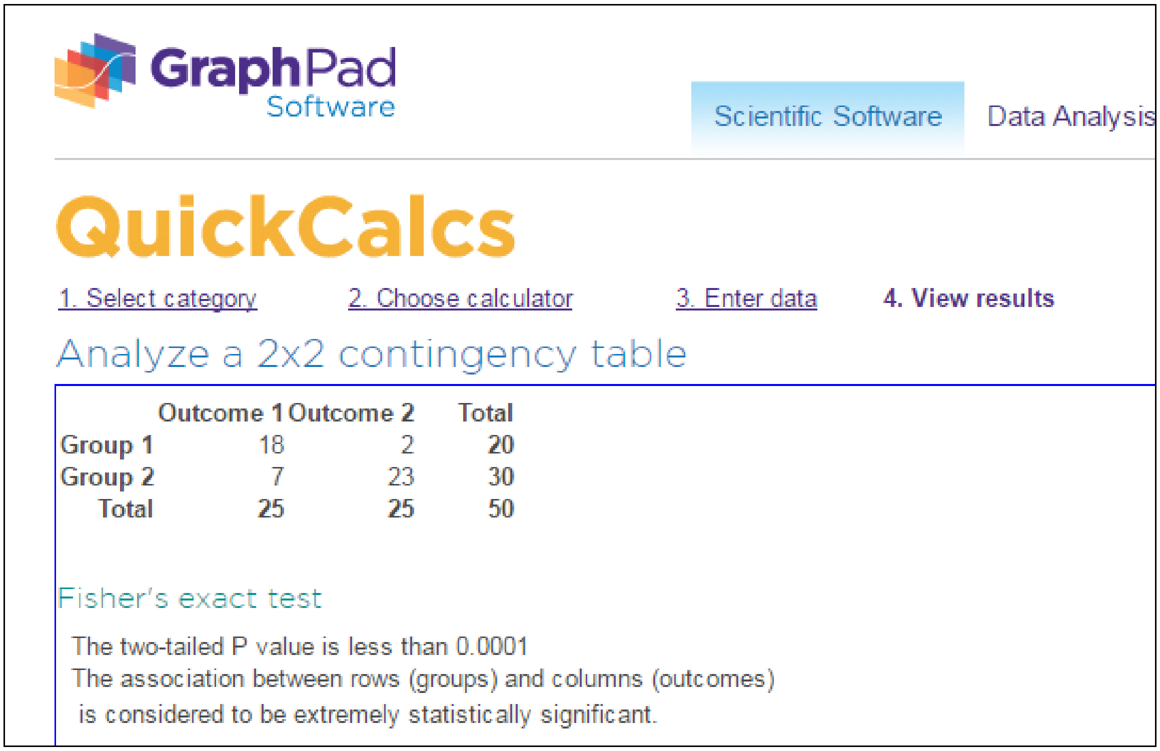
Cell D represents the letters that had been addressed to the Mosque, but did not arrive

Especially if the group sizes had been unequal (e.g., if A + B had not been equal to C + D), we should express these differences in terms of percents. The percent of Methodist letters delivered would be 100 X A / (A+B). The percent of Mosque letters delivered would be 100 X C / (C+D).

These results could be visually displayed within the contingency table

Contingency Table for Trace Field Count			
	DV = letter delivered	*DV = letter not delivered*	*Totals*
IV = addressed to Methodist Church	18	2	20
IV = addressed to a Mosque	7	23	30
Totals	25	25	N = 50

So, 90% of the Methodist letters were delivered, but only 23% of the Mosque letters were delivered. Significance would be tested by the *Fisher Exact* probability test and the results indicate excellent significance.



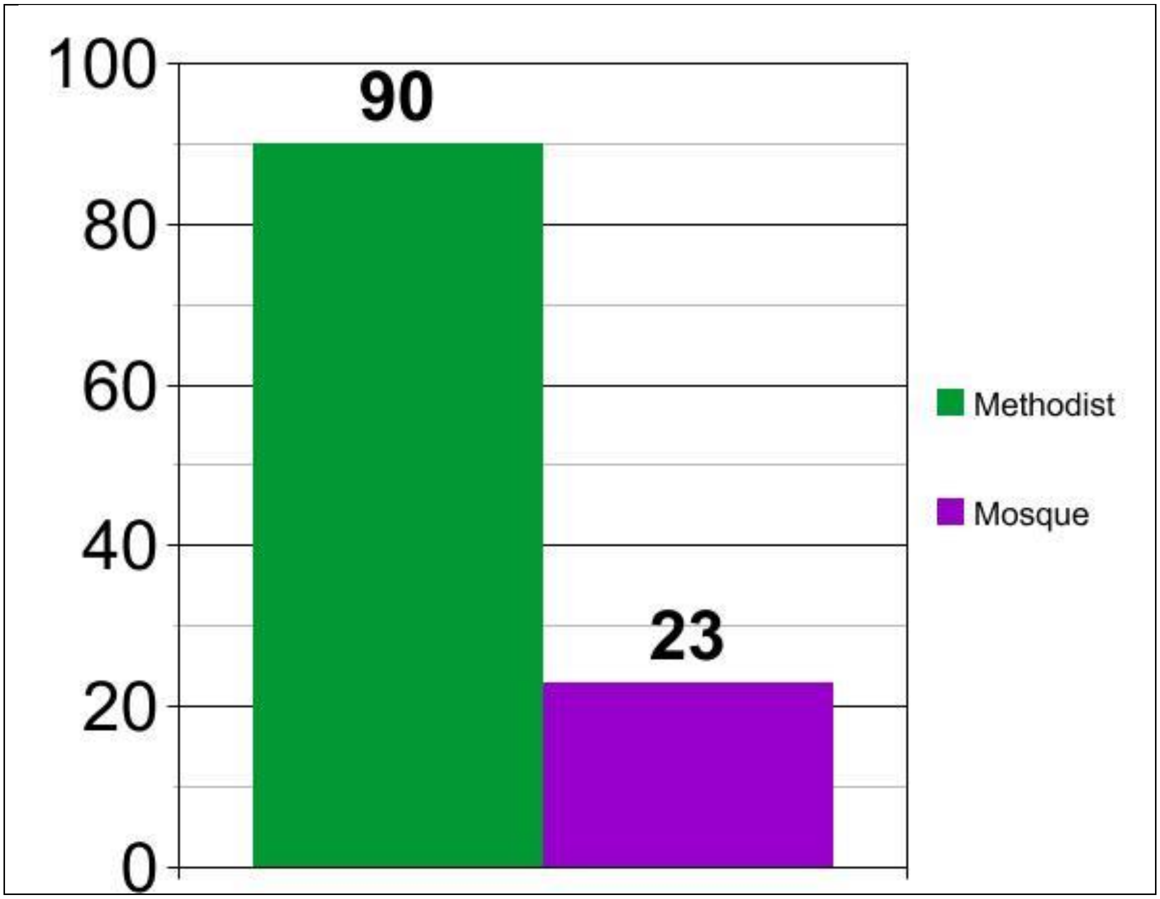
Analyze a 2x2 contingency table

	Outcome 1	Outcome 2	Total
Group 1	18	2	20
Group 2	7	23	30
Total	25	25	50

Fisher's exact test

The two-tailed P value is less than 0.0001
The association between rows (groups) and columns (outcomes)
is considered to be extremely statistically significant.

The bar chart could summarize the results.



A 21st century version of this approach might be to post a similar question on two different internet forums and see how many responses (and what kind of responses) each elicits, or perhaps post two different questions on the same forum and note the difference in response rate.

Or, you could make this an archival study by finding out which topics have been addressed by which forums. Just look at a specialized search site, like boardreader.

Other tracelike examples would be to count how many times a given tweet is retweeted, or how many hits a Youtube video gets, or how many “likes” something gets on Facebook.

But remember, you are not observing any subjects, only results that evidence that some subjects responded. (Indeed, with the build-in anonymity of the internet and fake handles, it is possible that just a few subjects are responding over and over again). You are definitely not able to measure how many people did *not* respond to each stimulus, so you cannot use a separate groups design with a two-by-two contingency table. You will have to use a sample versus norms design which assumes that in a random world, half of the responses would go to each stimulus.

Aggregates

Another approach to sampling, found mostly among sociologists, historians, political scientists, epidemiologists and economists is to have a sample be of *aggregates* of human subjects. For these we do use a rows and columns spreadsheet, but now, each row of data on the spreadsheet might come from an entire school, city, state, nation, hospital, company, or time period. These are now the cases, not the individual organisms.

Some of these examples are like experiments in that large units are treated differently. In 2000, the northern Mexican state of Coahuila provided free concrete floors for all homes in urban areas, including one of its largest cities, Torreon. Just across the state line in Durango was a comparable city, almost as large, Gomez Palacio. Two years later, various criterion variables were compared. Measurements were taken of parasite infection, anemia, school children's cognitive abilities, rates of adult depression and life satisfaction. On most of these measures, the average Torreon resident had improved and was significantly better off than the average resident of Gomez Palacio. The confounding variables would be any other differences between these two cities, either pre-existing (e.g., demographics) and different political administrations at the state and municipal level. Another possible confounding variable is that families who were more economically able (or just more concerned about health) moved to Torreon.

Another example, let's look at national statistics for two dozen industrialized nations (each nation on a different row of the spreadsheet). One column might be a rating of the comprehensiveness of that country's sex education and the other column might be the adolescent pregnancy rate. We would hypothesize a negative correlation between the amount of sex education in a country and its teen pregnancy rate (e.g., better sex ed in Israel, more pregnancies in the U.S.).

The limitation of this approach is that it aggregates all the data for large numbers of people, and can only tell us about the "average" Israeli teen and the "average" American teen and would not be able to tell us which kinds of teens are getting pregnant within each society.

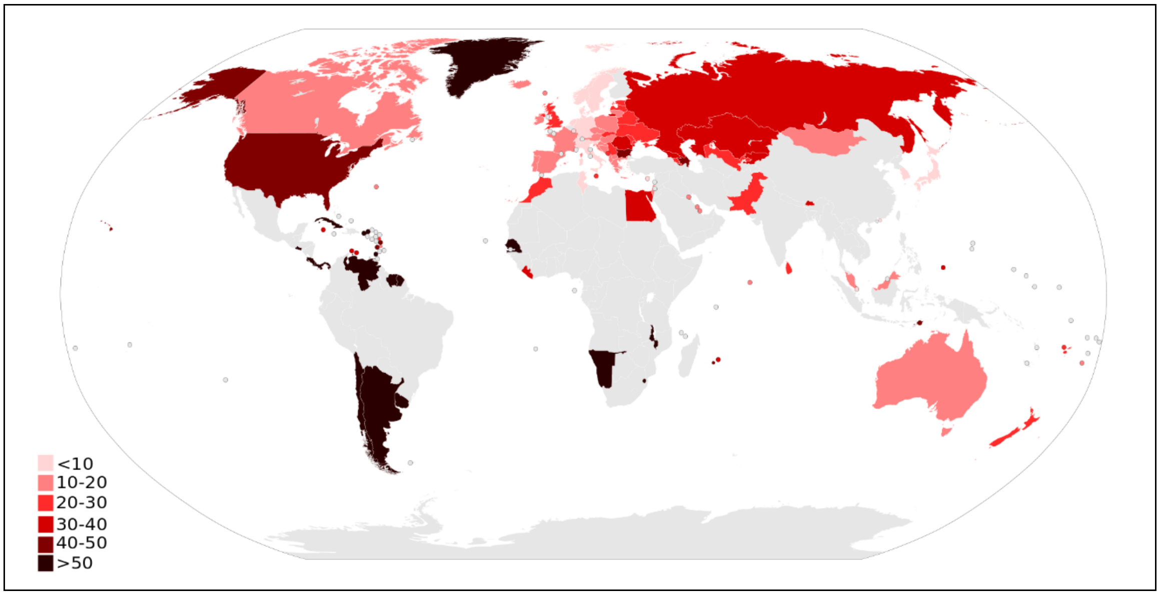
Another problem with these designs is the presence of numerous confounding variables. If we are comparing two time periods, before and after a given event, there were many things that changed during that period, and it is difficult to attribute measured outcomes to any one of them.

For example, did the introduction of ultrasound screening of pregnant women increase abortion rates? The rationale would be that if couples prefer a son instead of a daughter, ultrasounds would lead to the practice of selective abortion (terminating a female pregnancy so that the couple can try again for a male). In Haryana state in northwest India, the hypothesized trend has taken place over the past four decades. Back in 1981, before the introduction of ultrasound, the gender ratio was 108 males per 100 females born. In just twenty years, in 2001, the ratio became 124 males per 100 females born. The sonography scenario is plausible, but perhaps some other factor has shifted this ratio.

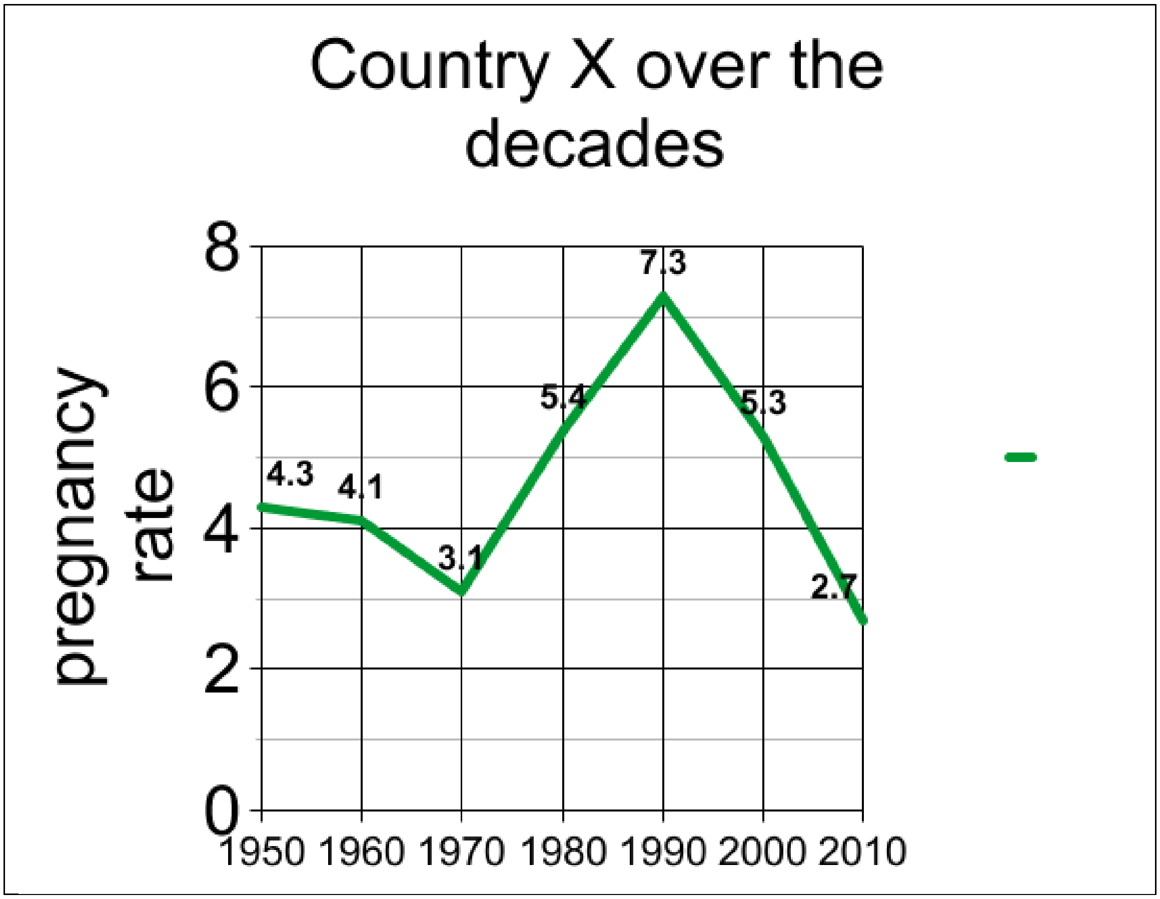
Another example of time differences of aggregates comes from Mexico. At the beginning of this century, a new administration was worried about the rising figures of obesity, rivaling those of its neighbor to the north. The federal government imposed a high tax on sugary soft drinks (e.g., Pepsi Cola). A year later, consumption of these drinks had fallen 12%, and has remained low per-capita. The decline was greatest among the poor, where these beverages were quite popular, and for whom the additional tax would have been most burdensome.

The descriptive statistics, inferential statistics, and visual reporting for these aggregate studies depends upon the design and the respective scalings. So, binary nominal scaling would be the two-by-two contingency table, percents, and Fisher Exact. If we had ratio scaling, we would use means and standard deviations for each variable, correlation coefficients, and a scatterplot.

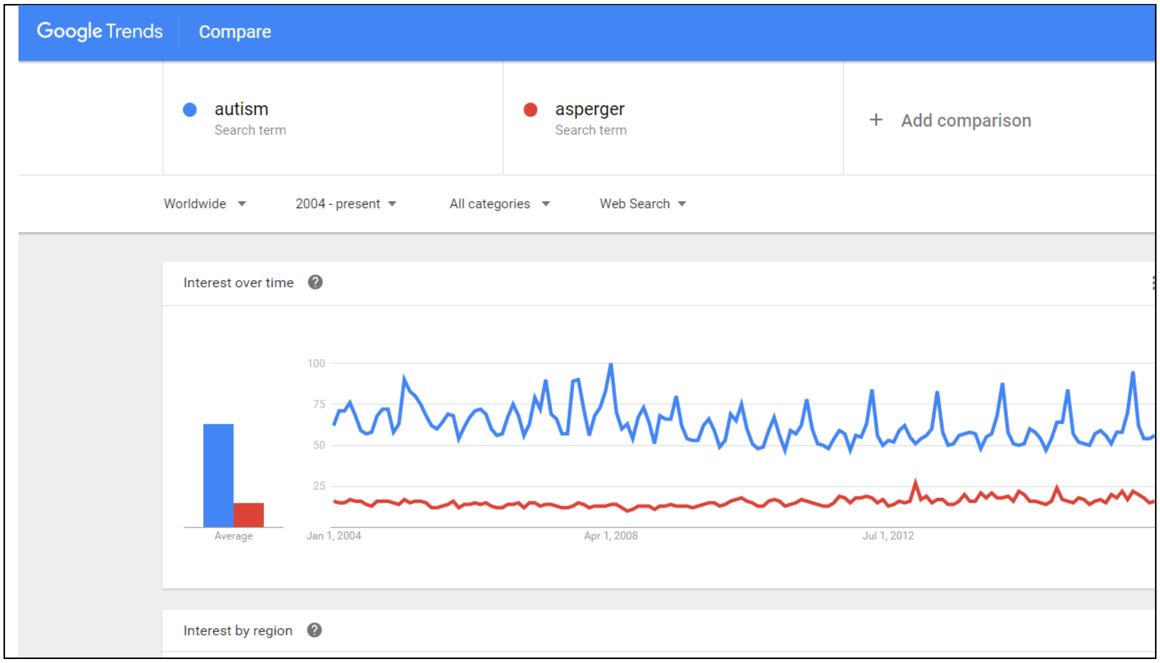
When the aggregates represent different locations (e.g., hospitals, cities, states, countries, regions) a map is a useful infographic.



When the aggregate represents different time periods, a line graph is most appropriate.



Another approach to measuring frequency of traces, rather than specific human subjects, comes from looking at electronic tracks. *Google Trends* notices how often a search term was used. A study can compare two different search terms (e.g. autism & Asperger), two different time periods (e.g., before and after the introduction of DSM-5), or two different geographical locations for the number of searches coming from those areas (e.g., Canada & Australia). Two different terms can be simultaneously tracked over time (i.e., week by week) and their relatively frequencies correlated. Best of all, Google Trends assembles the data and downloads to a CSV file (which can be uploaded into Excel, Statcato, or JASP.

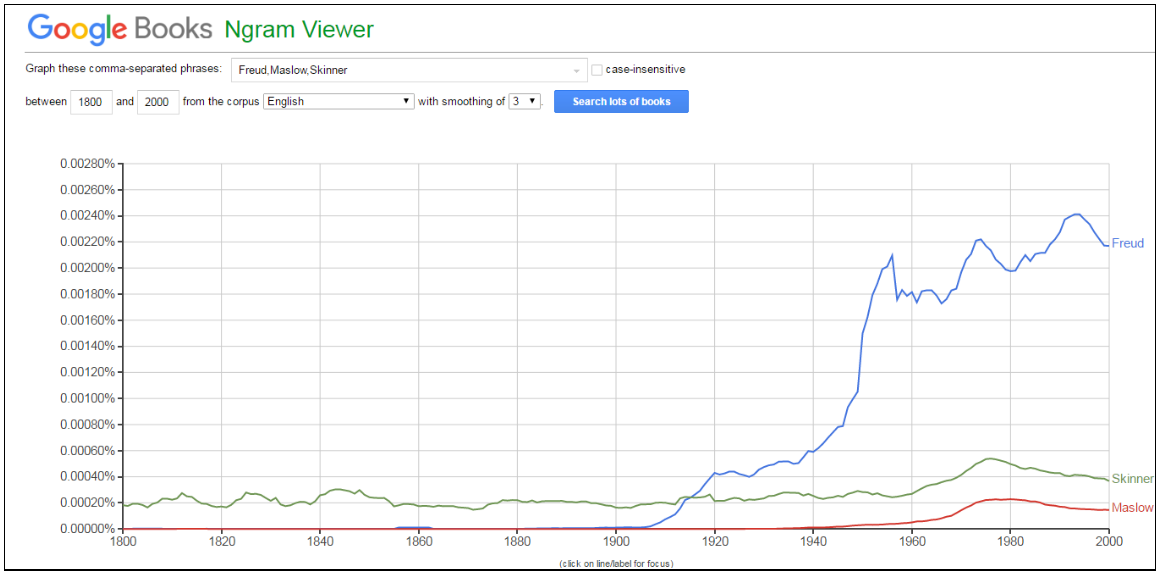


● autism
Search term

● asperger
Search term

Worldwide ▾ 2004 - present ▾ All categories ▾ Web Search ▾

The frequencies of the appearance of specific words can also be tracked in other databases, such as ngrams or bibliomania. These would allow us to compare the frequency of the names Freud, Skinner, and Maslow in books. (The latter two peaked in the late 1970s, but Freud continues to draw comment.) A confounding variable would be other persons with that name.



Graph these comma-separated phrases: Freud,Maslow,Skinner □ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3.

You could do an advanced search to compare different time periods and languages of books in terms of their mention of Freud.

Language	1900-1909	1910-1919	1920-1929	1930-1939
English	13	17	16	22
German	15	13	20	13
French	11	14	16	27
Spanish	1	4	16	16

Obviously, these book searches do not constitute looking through the population of all books published in a certain time period, language, or geographical region. The sample is small and there is no indication that it is in anyway randomly selected or representative.

Many professional journals now offer great online search capacities for specific words and phrases. This could be useful not only in your literature review, but also in tracking how popular a topic was in different periods of history.

Single Subject

Among clinical psychologists, especially behaviorists, a popular technique is the *single subject design* (also known as *small n*). Similar techniques were used in psychology's earlier history by Wundt, Ebbinghaus, Pavlov, Watson and Skinner. It is used today, especially with Applied Behavior Analysis. The idea is to get an interrupted time series: composed of an *ABA* design: before treatment (for a baseline or control condition), during treatment (hopefully showing favorable impact of the treatment), and then after the treatment has been withdrawn (and the expectation is that the organism will regress to pre-treatment levels). This last phase of measurement does not apply if we expect the treatment to be a permanent "cure" rather than a "maintenance."

This may look very much like a repeated measures design (and it does share much of the sequencing weaknesses of such a design). However, we cannot run those types of inferential statistics on n = 1 designs. What we can do would be to treat all the data measurements for a variable under condition A as one group, and all the data measurements under condition B as another group, and then run separate groups inferential statistics.

Such small n designs may have high *internal validity* (because there is not a great variation between so few subjects). However, *external validity* is low, because we cannot assume that the small sample could be representative of the entire population: what has worked in this one case may not work in all, or even most patients.

Big Data

What does the future hold? The old paper and pencil questionnaire will be gone. We can get those data (and a lot more) through continuous monitoring of individuals' behavior or their electronic activity. Here are the benefits of using the stream of information provided by an individual's smart phone communication, web search, fitbit and real-time laboratory data. Such data will be

- Continuous flows of updates rather than one snapshot of the individual at a given time

- Precisely measured by physical, chemical and biological processes rather than the individual's attempt to put subjective physiological and mental experiences into words

- Objectively reported rather than filtered by the individual's concerns for self-presentation

- From a larger and more representative sample because it will be harder to opt out

- Automatically coded into spreadsheets or other formats for statistical analysis

- Easily transferable between researchers in order to facilitate meta-analytic studies

Chapter #13: Qualitative Methods

The essence of *qualitative* methods is that data are more *rich* when in a *narrative* form. In practice, this means respecting the subject's own words and attempting to empathically comprehend the meaning of those words for the subject. This table shows the range of human knowledge provided by different levels of research, from the extreme of the most rich (but least precise) to the other extreme of the most precise (but lacking in richness). The qualitative research discussed in this chapter represents the range between the metaphorical representation of reality and the numerical representation of reality. At this level, reality (or its subjective perception) is represented by words and images that have meanings, and the task of the qualitative researcher is to understand those meanings.

Most rich	Mystical experience: all words, numbers and contact with material world is transcended
	Metaphoric: the realm of art, music, ritual, poetry
QUALITATIVE	Narrative: subjective account of individual experience
QUANTITATIVE	Binary Nominal: dichotomous categories of (e.g., yes/no, pass/fail, experimental/control)
	Multiple Nominal: more than two categories
	Ordinal: ranked levels (e.g., excellent/good/fair/poor)
	Interval / Ratio Discrete: whole numbers represent quantities (e.g., incidents, units produced or sold)
Most precise	Ratio Continuous: numbers can be in fractions and decimals

It should be noted that some statistics textbooks use the term qualitative to include the nominal levels of the above diagram. To reiterate, in this course, *qualitative* only refers to the narrative level of data, depicted above in light green.

Many scientists, even psychological researchers, are skeptical of these narrative data. Most people are convinced of the unitary nature of truth, such that once they have come to trust one avenue for getting at the truth, they tend to distrust all others.

Perhaps this would be a good time to review some of the classic studies of our science. They are best remembered for the words of the participants and images of their behaviors. Harlow's study of motherless monkeys concluded that infants need nurturing. Zimbardo's Stanford Prison Study showed that decent, normal people could be corrupted by institutional roles. Milgram's study of obedience demonstrated that most people would follow the commands of an authority figure and punish a stranger. Clark's black doll study found that racist norms had been internalized by African-American children. Qualitative research is more vivid and illustrative, and that makes it more interesting to do and inspiring to read about. No improved statistical significance could make these classic studies more profound than the face of the little girl who selects the white doll.

Great theories from Maslow's levels of needs to Piaget's and Erikson's stages of development were built upon case studies, not statistical analyses of tests with cut-off scores. Of course, none of these qualitative studies should be viewed as closing the book on research in these areas. Indeed, if we look at the qualitative research generating the theories of Maslow, Erikson, Piaget or even Freud, that should inspire us to develop quantitative studies (surveys and experiments) to confirm the insights of those theories.

	Quantitative	**Qualitative**
Focus on	Outcomes	Process
Requires	Objectivity	Disciplined subjectivity
Goal	Discover causal relations	Explore meanings
Offers	Precision	Richness
Subjects are	Lab animals or people who filled out questionnaires, left traces, or files in an archive	Patients, historical figures, informants about cultures, organizations, creators of writings, works of art
Data scale	Nominal, ordinal, interval or ratio	Narrative or visual
Introspection regarded as	Unscientific	Essential
Sample size should be	Large enough to randomize inter-subject differences	Small enough to explore intra-subject meanings in depth
Hypotheses are	Tested and confirmed	Generated
Researcher's virtue	Dispassionate, detached	Engaged, builds rapport
Sensitive topics require	Anonymity	Empathy, trust, rapport
Data come from	Archives, field counts, questionnaires, traces, laboratories	Introspection, case studies, ethnographies, field studies, participant observations, interviews, focus groups, text analysis, visuals
Data are coded into	Numbers	Patterns
Reliability means	Consistency of data as demonstrated by strong correlations	Trustworthiness or credibility of data as demonstrated by saturation and triangulation

	Quantitative	**Qualitative**
Designs	Sample vs. Norms, Separate Groups, Repeated Measures, Correlational (these four are clearly distinguished)	Phenomenological, Grounded Theory, Content Analysis (these blend into each other)
Conceiving the design	Challenging	Easy
Gathering data	Tedious	Fascinating
Coding the data	Tedious	Challenging
Write up	Easy but formulaic: go through the hypotheses	Very challenging: tell the story
Questions result in answers that	Become scored responses	Elicit deeper reflection and dialogue
Greatest insult you can give to a researcher using this	"You are just a bean counter, measuring what was easiest to measure, not what was most important to really know."	"You are just a hack, writing about what was most emotional."

Qualitative research focuses on the subject's own words.

Introspection

Before psychology became established as the scientific study of behavior, it was known as the study of the mind. Its first research technique was mere *introspection:* the researcher would simply reflect on his/her own thoughts, emotions and actions, asking "Why do I think what I think, feel what I feel, do what I do"? Many of the great pioneers of modern psychology (e.g., Wilhelm Wundt, William James, Mary Calkins) were primarily doing introspection. What distinguishes their introspection from the kind of self-reports that participants in a survey use to answer dependent variable questions is that the former has no distinction between the subject and the researcher: the observer is the observed.

It was John Watson whose criticism of introspection as inherently unscientific redirected psychology to try to be more of a laboratory science, even if that meant that more research was to be done on caged animals rather than thinking humans. Watson was right that introspection never proves anything in the sense of being able to confirm a specific causal hypothesis.

One problem with introspection is that the sample size (n = 1) is too small, and probably not representative of the population. There can be no inter-rater reliability, because you cannot explore my thoughts, and I cannot get into your mind to explore yours. Each of us is limited to the exploration of our own thoughts.

Another problem is the lack of distinction between the researcher and the research subject, between the datum and the analysis of that datum. When I am thinking about my own thinking, there is no clear guideline as to where the data of thought ends and where the interpretation of those data begins. When is the psychologist the observed, and when the observer?

Then there is the problem of bias: the researcher may want to present self in the best possible light, and perhaps as more full of socially approved motives and thoughts than is actually the case. (Freud's courage was his acknowledgement of his own incestuous and parricidal urges.)

Introspection is never an adequate approach to research, but it is both unavoidable and essential, a starting point for any cognitive, affective, or behavioral exploration, and a constant guide to prevent us from infusing our own data into the experiences of our subjects. We had better be doing more than mere introspection, but we had better be doing introspection in order to limit its impact on our own interpretations.

Unless the topic of your research is purely physiological or unless your subjects are non-human, begin your research with introspection. This process can generate a hypothesis, suggest how to obtain participants, and refine the wording of a questionnaire. Return to introspection after your data have been analyzed statistically. Now introspection will suggest some underlying causal connections and future paths for the next step of research.

For example, let’s suppose you want to develop a research topic within the branch of psychology that studies the marketplace, the field of consumer behavior. A friend is giving you a ride back from the airport and wants to stop off at a discount grocery store in San Bernardino. You have never been inside that store, or even to that part of town, but at your friend’s urging, you go inside. You see a sale on your favorite snack bars at an incredible price. They are stacked on pallets, still in large boxes with Chinese markings, and you decide not to get any. From a purely quantitative framework, we would record your visit as follows. “Did you purchase anything? yes / no” (answer no). “How many power bars did you purchase?” (answer 0).

This is where the value of introspection comes in. It goes beyond the yes/no and the how much. It can ask why. You didn’t make a purchase because of some stimulus, now, just what was it? Was there a bad smell in the store? Did you start to wonder if the food had really been made in China and the packaging was counterfeit? Did you fear that the packages had been sent over the ocean to China, and were not able to pass customs there and had to be returned to the U.S. (and are now stale)? You have now come up with some hypotheses and three different independent variables to manipulate in future experiments.

Just remember this: you are enough like other people so that you should not assume that your own thoughts, emotions, and responses are incapable in other individuals. But you are unique in enough important ways that you should not assume that everyone else is always going to have the same experiences in the same situations. So, we allow introspection to generate a hypothesis, but not provide data for its confirmation.

Case Studies & Ethnographies

The other great research technique of the pioneers of modern psychology was the *case study*, which attempts to do an in-depth study of a single subject. There are two main forms of case study. The *biography* (or life history) is as old as the writing of history. In most cases, the researcher has never met, seen, or spoken to the subject, but must rely upon an analysis of extant documents. Ideally, this would include an autobiography, diary, personal letters, speeches, other literary and artistic productions by the subject, photographs, videos, and writings of contemporaries.

The other form of case study is the clinical, which includes narrative (as well as quantifiable) data from interviews, testing and the response to treatments (e.g., medication, psychotherapy). These data can be as qualitative as free association and dream analysis, and as quantitative as laboratory results and psychometric scores. If there were several sequential measures of these quantitative variables during a before treatment / during treatment / after treatment period, that could qualify as the ABA design of single subject quantitative research.

Case studies have many of the same limitations as do introspection. The sample size is too small (n = 1) and the sample is not representative, especially since the cases are not selected because they are typical, but because they are unusual and challenging. Biographies are written of famous people whose accomplishments are extraordinary.

Bias can be a problem for case studies because researchers may adhere to a particular school or theory (e.g., psychoanalysis) and, at least subjectively, may hope that this case will illustrate. The *American Journal of Psychoanalysis* does not publish case studies in order to demonstrate the inadequacy of Freudian technique, but to illustrate how it can be applied to new conditions.

The focus on the illustrative case study has been maintained in other social sciences (e.g., history, anthropology). In the latter, this is in the form of an *ethnography* where the subject is not an individual person, but a cultural system. In organizational studies, the field work might be in the form of *participant observation* in which the researcher acknowledges the impact of her interactive behavior on the system being studied.

A great site to do a participant observation study or ethnography (and log some service learning hours) would be some agency (government or private charity) providing some kind of social service. These sites could focus on the motives and behaviors of the recipients, the working staff or the volunteers, and are great examples of human interaction and organizational culture.

<table>
<tr><td>Suppose you were using participant observation at one of these rallies. Could you conceal your own values? Could you report your findings in an objective fashion?</td><td>
</td></tr>
<tr><td colspan="2">
</td></tr>
</table>

Interviews & Focus Groups

The greatest interactive feature of the clinical case study is the *interview*, both in the form of the intake questions about the patient's presenting problem and the empathic probes used within psychotherapy. Unlike the kinds of *closed-end*, easily quantified responses associated with the questions appearing on a questionnaire, the true interview has *open-*

ended questions permitting a variety of responses (many of which cannot be easily scored on nominal, ordinal, interval or ratio scales). Indeed, the investigator may not know what answer, or even what type of answer, the subject might come up with.

This video demonstrates this branching approach, used in a life history context with time anchors.

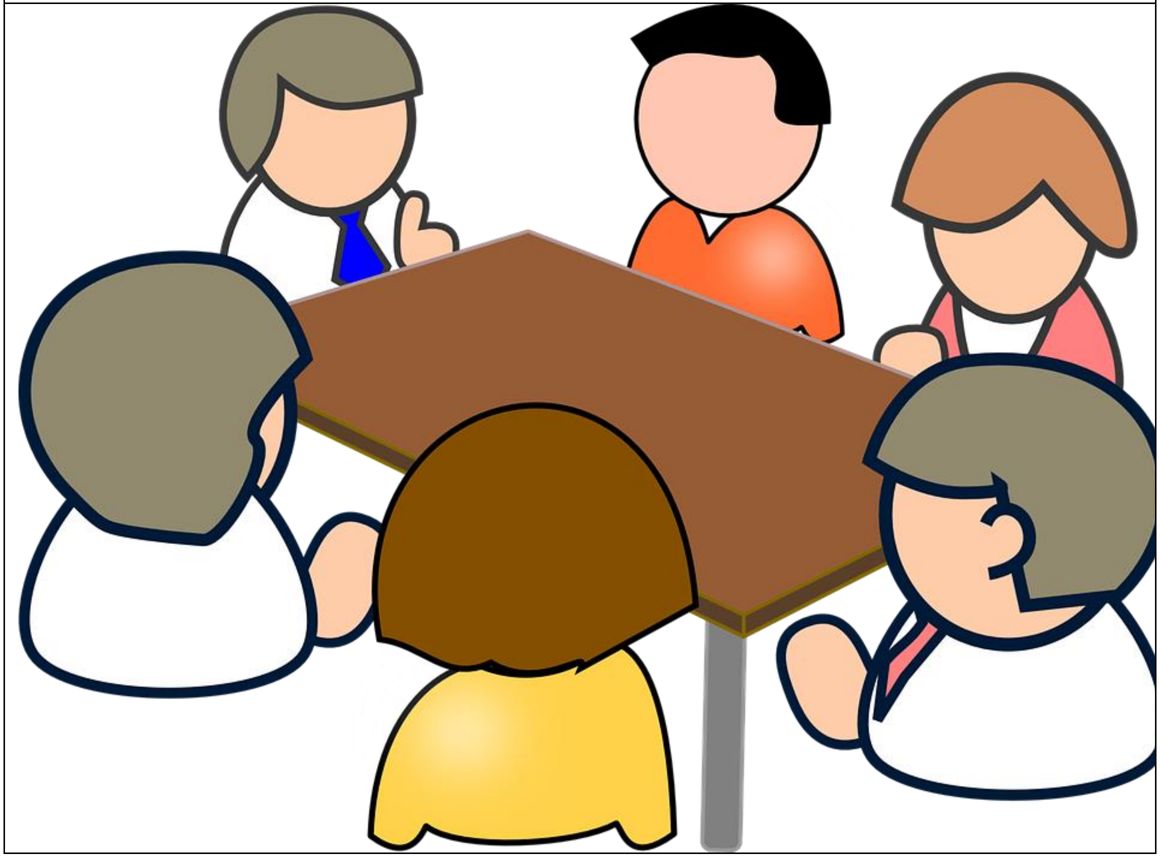
While questionnaires can be conducted with pre-printed sheets of paper containing all possible alternative answers, and automated websites with drop down menus for answers, interviews do not work well with these restrictive formats. Just leaving a blank space on a piece of paper, after a printed question, will result in poor quality narrative responses. The rule of thumb for a quality interview is that the researcher should be synchronously interacting with the subject, and that the optimal response formal is oral rather than typing or writing the answer. Typing the answer does not work whether it is on a sheet of paper or email or tweets. Longer typewritten answers (such as a threaded discussion, journal or blog, may fit better into the type of textual analysis given below).

	Questionnaire	Interview
Research	Quantitative	Qualitative
Interaction with researcher	Can be Asynchronous	Should be Synchronous
Questions with limited, quantifiable answers	Excellent approach	Poor approach
Printed questions with just a space to write an answer	Poor approach	Poor approach
Orally presented questions	Can work if answers are codable as numbers	Necessary for rich answers

One of the most effective forms of interviews used in qualitative marketing research is the *focus group*. Here there are between four and fifteen participants who synchronously interact with the research and with each other and orally answer open-ended questions (usually about a product or service). The exact questions and their sequence may be arranged extemporaneously. A particular answer (especially if somewhat unexpected) may trigger the examiner to follow along a different line. The goal is to understand the decision-making process of the consumer, why certain features lead to a product being excluded from further consideration as a viable alternative. I have another book (also a free download) about marketing research.

This group interview serves to create a critical mass for interaction and reflection about the product or service. Each consumer's comments elicit new ideas in the other participants.

Focus groups can be used by product developers or by non-profit agencies to discuss service delivery.



Content Analysis

Scholars who write history or literary analysis also use *content analysis* of text documents in order to search for identifiable themes and patterns that may provide a key to underlying meanings. These text documents can be the subject’s journals or diaries or transcripts of conversations. Some of the easiest data to access for content analysis in the 21st century would be from the internet: email, tweets, Facebook updates, and posts on discussion boards.

Fortunately, there are also some easy to use tools for tracking these, such as boardreader.

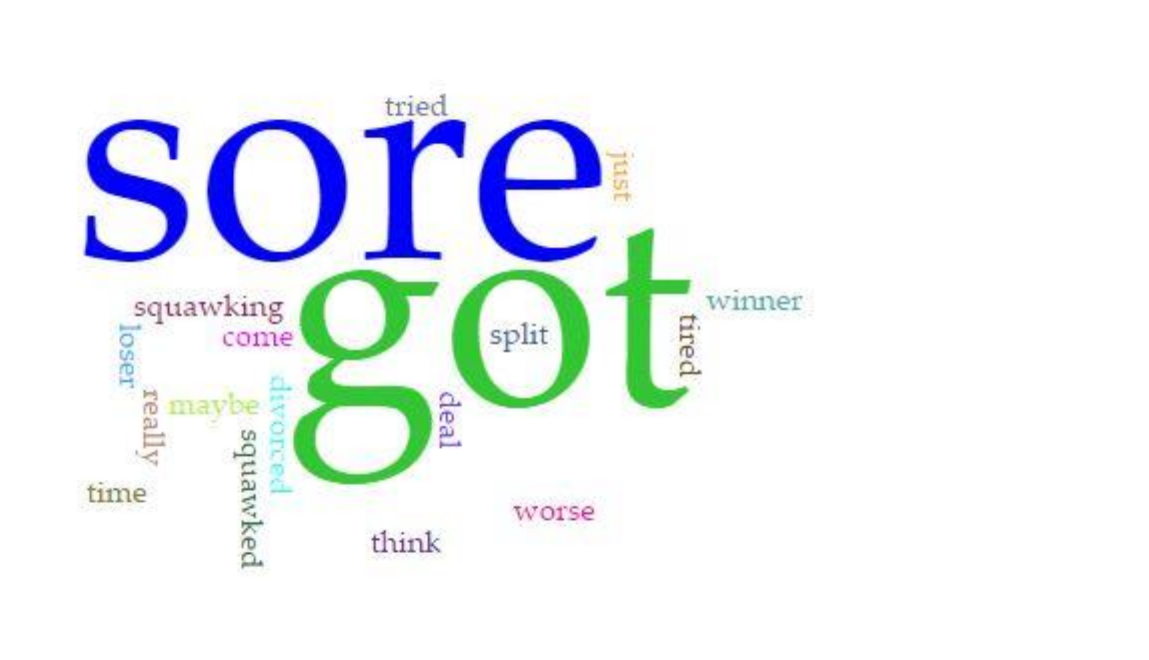
One big challenge comes in defining the sample. Sometimes we cannot tell if one person is just one “handle” or several. Some people have a large volume of writing on social media while others post, update and tweet sparingly. We have to decide whom to sample and which of those person’s words.

Then the challenge is to make sense out of all of these words. There are numerous computer programs for such analysis. Some are expensive and most require quite a bit of training in order to understand how to use the features (and apply some of the linguistic theories which underlie them). Here is some freeware. This Visual Understanding Environment VUE is from Tufts University requires some training.

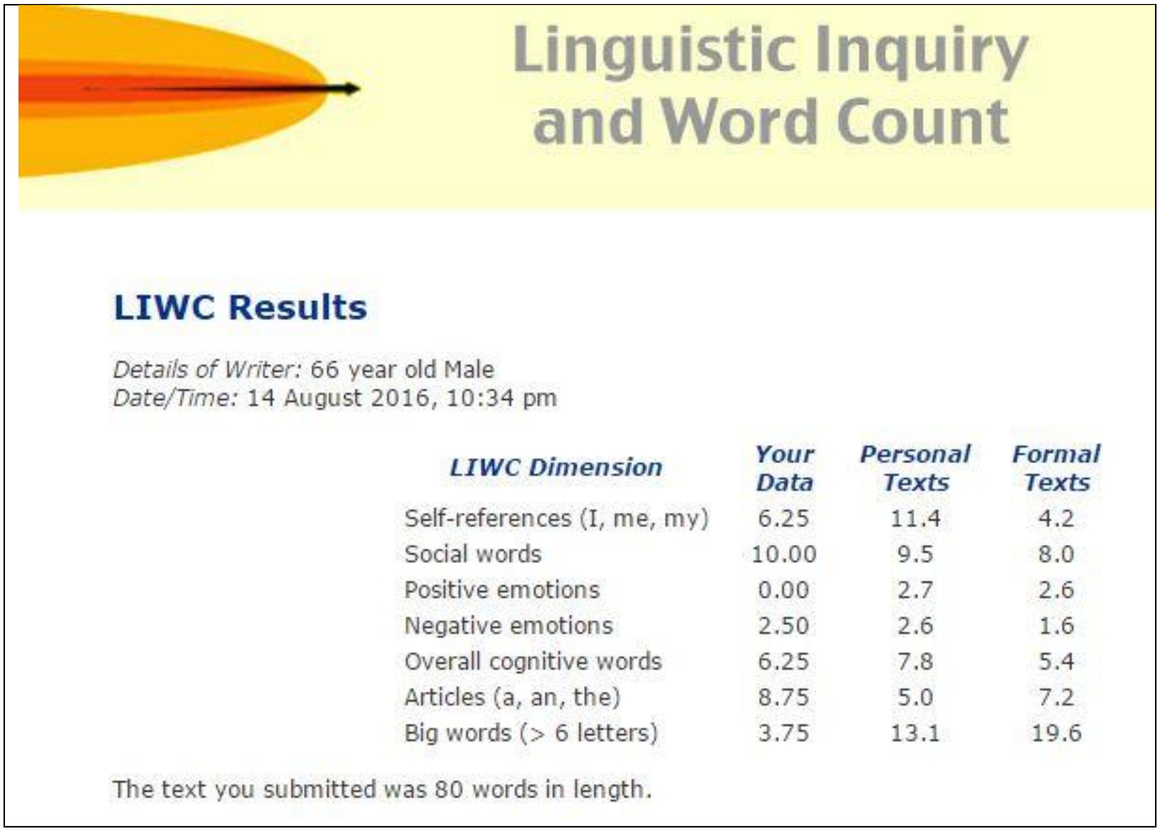
To see how some of the easier sites work, let’s take a passage of text that might appear in a subject’s diary, private letter or psychotherapy transcript.

“I divorced her (not the other way around). I just got tired of the deal. I could put up with an 80/20 split, or maybe even a 90/10, but it was her way 95% of the time. And that 5% where I tried to get my way, she really squawked about it. Come to think of it, she would keep squawking even when she got her way. She was worse than a sore loser. She was a sore winner.”

Voyant is one of the easiest to use. It makes nice word clusters.



Some of the more elaborate content analysis programs attempt to provide some quantification of words and phrases used. This is the Linguistic Inquiry and Word Count LIWC.



Linguistic Inquiry and Word Count

LIWC Results

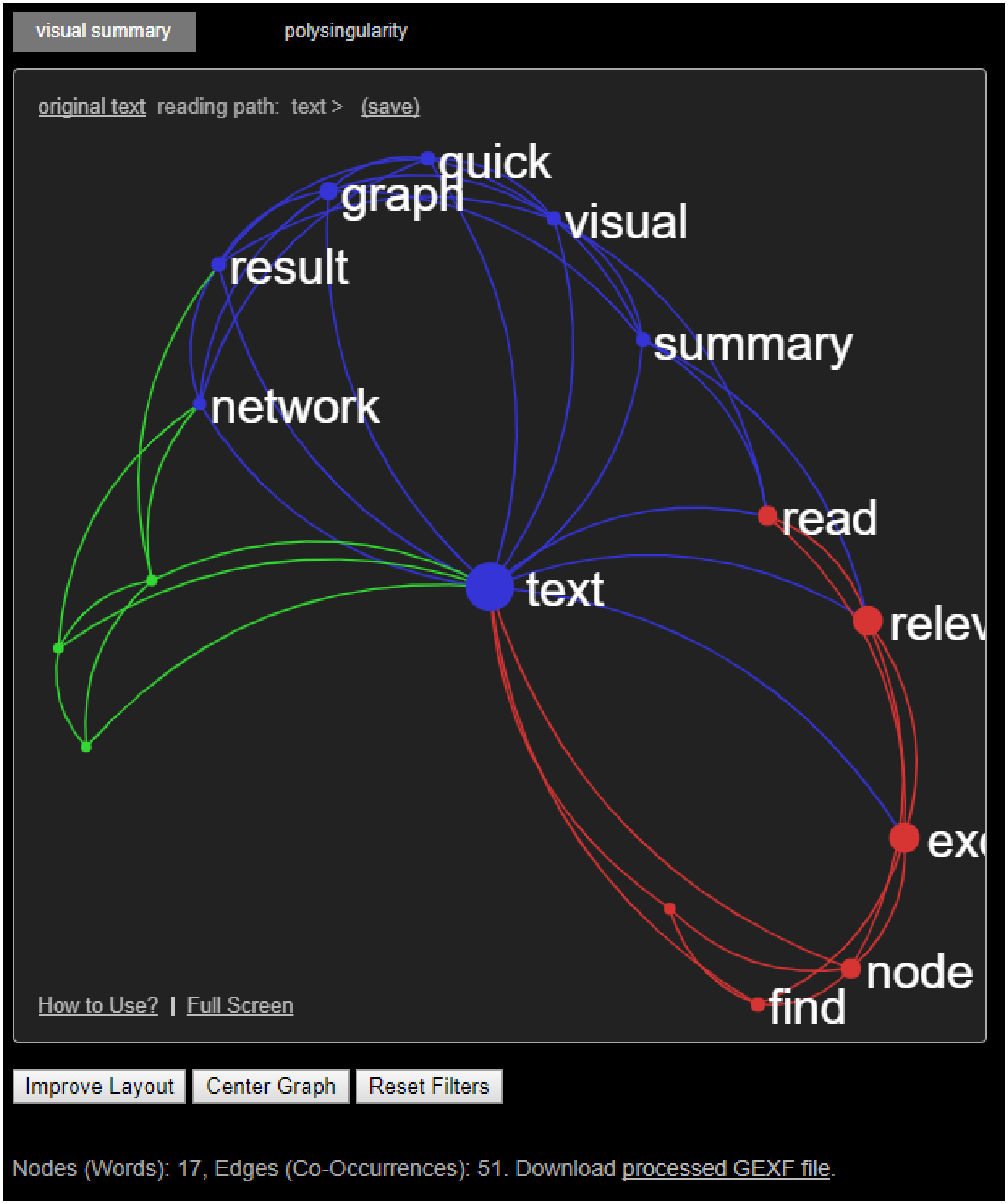
Details of Writer: 66 year old Male
Date/Time: 14 August 2016, 10:34 pm

LIWC Dimension	*Your Data*	*Personal Texts*	*Formal Texts*
Self-references (I, me, my)	6.25	11.4	4.2
Social words	10.00	9.5	8.0
Positive emotions	0.00	2.7	2.6
Negative emotions	2.50	2.6	1.6
Overall cognitive words	6.25	7.8	5.4
Articles (a, an, the)	8.75	5.0	7.2
Big words (> 6 letters)	3.75	13.1	19.6

The text you submitted was 80 words in length.

Notice that the results (going against the conventions of clinical psychology) use “positive” to imply good and “negative” to imply bad. The use of the above numbers in interpreting the passage of text would be that it correctly identifies these words as something said when the individual was experiencing a low mood, a combination of sadness and anger. This passage is heavy on the social words but very low on the big words: it is a heart-felt explanation of a failed relationship.

This next site is called textexture. It requires you to register, but has the coolest diagrams for connecting words.



Nodes (Words): 17, Edges (Co-Occurrences): 51. Download processed GEXF file.

Art historians can do a study of visual images (e.g., photographs), but this will require even more subjective interpretation on the part of the investigator.

Sampling

Sample size in qualitative research can be smaller since we don’t need to worry about getting enough subjects to attain statistical significance. Rather than strive for a representative sample that has not been self-selected, qualitative research looks for those informants who are most willing to disclose personal information and can do so in an articulate manner. Rather than lining up all of our informants before the beginning of data collection, a frequently used approach is a *snowball* technique in which we develop rapport interviewing one informant, who then refers us to other good informants.

Knowing when to stop data collection can be determined by several factors. One is looking for *saturation*, when additional informants seem to be repeating what we have already absorbed from previous informants.

Conclusion

In all of the above qualitative techniques, the researcher must go beyond being a detached and purely objective recorder of easily measured data. The researcher must engage the subject, building rapport, eliciting an in-depth response. The researcher than introspects to co-create an analysis of *connotative*, rather than purely denotative meanings, using introspection to elucidate what the informant must have experienced (rather than distort what the subject's own meanings are). While quantitative research is designed to learn about a topic, qualitative research facilitates the process of learning with. Research is less of a mechanical data gathering, and more of an ongoing interpersonal relationship. A good follow-up is to debrief the subject, and rethink the theory, if it does not fit the subject's experience. Through everything, we are learning more about our subject (or the site of the investigation) than we are of the topic per se.

The standards of validity and reliability are clear in quantitative research, and can be precisely expressed by correlation coefficients (e.g., Pearson, Spearman, Cronbach, Kuder-Richardson). In qualitative research these concepts of validity and reliability need to be reconceptualized. The narrative data must be trustworthy or dependable in the sense of being what the informant really feels and thinks, rather than what he says in hopes that it will be what the researcher wants to hear. Here the virtue of empathy and rapport are necessary, especially on sensitive topics (such as sex, addictions and intimate violence). We achieve something like inter-rater reliability when subsequent informants seem to be repeating what we have already heard from previous informants, and this is known as *saturation*.

Compared to quantitative research, the qualitative is much easier to formulate a topic, and the data gathering is more fun. The data coding won't be tedious, but transcribing oral interviews can be time consuming. The greatest challenge comes in the write up. There is no major formula, just vague guidelines like “tell the story” or “let your theory be grounded in the data.” Every other quantitative study you read makes you more certain about how to write up your own quantitative study, but every time you read someone else's qualitative report, the less certain you become of how to write up your own qualitative study. So, if you want to finish your dissertation quickly, keep it purely quantitative and wait until you have more time before you commit yourself to doing a qualitative study. Just remember, although we often refer to qualitative research as “soft” that doesn’t mean that the write-up is easy.

The debate should not be whether psychology or any other social science should be purely quantitative or purely qualitative, but how can we do both better? The mark of well-done quantitative research is that we discover precisely what were the individual choices made by our subjects, and how these correlate with background factors and stimuli. The mark of well-done qualitative research is that we explore richly how our subjects make those choices.

There is an alternating cycle of research. We should usually start with the qualitative (at least with introspection): generating a hypothesis. Then we should switch to the quantitative, a correlational study or experiment in order to test that hypothesis. Then we go through another qualitative cycle (perhaps a focus group), trying to explain the reasons *why* behind the observed correlation, and these reasons become our new hypotheses. Then we do another quantitative, testing out those new hypotheses. This cycle never ends, and that means that scientific knowledge is never static, but always growing and perhaps even changing.

If you are not ready to do a completely qualitative study, just try adding a small qualitative component to your quantitative study. This small supplement might help your discussion section in which you attempt to explain your results. For example, in a field count or trace research, just introspect be aware of your own thoughts and expectations. Also, listen to the conversation of your subjects about why they do what they do. If you are observing students cut across the grass in order to get to class (or traces of this: the pathways cut into the grass), what are students saying when they do this? What does their body language say? Is this a conscious decision made with some guilt or an automatic response?

If you are doing an archival study, is there a chance to get some images, or better yet, the words of the subjects (or those who evaluated the subjects) that have been entered into the records? Remember the Billion Graves project? Can you get some pictures? Can you read some inscriptions? Those qualitative data will tell you more about the meaning of those graves than you can ever get from sheer numbers.

Chapter #14: The Report

The basic format for reporting research is in a write-up that may serve as a term paper, senior *thesis*, doctoral *dissertation*, conference presentation, or article to be published in a *peer reviewed* journal.

Organization

One of my favorite trite tautologies is “Time is nature’s way of preventing everything from happening all at once.” That also applies to how we organize a poster space, oral presentation or write-up. We don’t want to say everything at once, and we don’t want to say it over and over (but we want to make sure that it does get said, preferably at the right time). So, whether the write-up is for a term paper in a class or an academic journal article, (or an oral presentation or a poster) there is a definite organization that has some prescribed sections to be covered. Each section is devoted to cover something specific, and should avoid covering other things belonging in a different section.

Title. The first thing the reader should see would be the title, although the final version of the title might be the very last thing that the author tweaks. For most journal articles, the title is a non-sentence arrangement of key words (usually emphasizing the criterion variable). The subtitle conveys more key words, perhaps indicating the predictor variables, manipulated independent variables, and/or population. Sometimes, especially in more popular journals, the title is presented as a question. For example, a good title for a journal dedicated to psychiatric research might be:

“Depression in Later Life: limitations of self-report scales”

while a more popular outlet might prefer something like

“How to know if Grandpa is depressed? Don’t trust the test!”

The next thing that the reader would see after the title might be the author name(s), institutional affiliation(s), contact information and declaration of support (e.g., from grants). If you are submitting your write-up for publication in a peer reviewed journal, the editorial guidelines may require you to submit such identifying information separately, so that the manuscript can be reviewed anonymously. In this way, no one could claim that the reviewers accepted your article just because you are so famous or popular (or are affiliated with such a prestigious institution), or that your article was rejected just because you are student at a small community college.

Abstract. The next thing the reader will see is the abstract. This is a summary of the article. This is written after the rest of the article has been completed. The only exception to this is that some conferences want you to submit an abstract, usually an elongated one, as a proposal of what you are going to say in your oral or poster presentation. In that case, the abstract is written after the data have been collected and statistically analyzed, but perhaps before the entire discussion section has been written.

SECTION	COVERAGE	WHEN TO WRITE IT
Title	State in terms of the criterion variable.	**9.** After the abstract.
Abstract	Summarize the entire write-up, emphasizing the methods & results sections.	**8**. After inferential statistics have been calculated and you have looked at the guidelines of the conference or journal.
Introduction	Review your topic: trace the development of the theories surrounding it and data from previous studies. Lead into hypotheses. Do not mention your own results yet.	**2**. Start after proposal is approved, keep adding references until it is due.
Hypotheses	Set these up as specific predictions, perhaps as part of the introduction. Do not mention whether your results confirmed them yet.	**1.** After proposal is approved
Method	Describe the site where research took place, the population or sample used, the operational definitions of each variable, the design by which hypotheses were tested, if some subjects were excluded, and data analysis. Do not describe the results for the criterion variables yet.	**4**. After data have been tabulated with descriptive statistics for all predictor variables. You can even start on parts of this section as soon as you have the apparatus (questionnaire)
Results	Present information on key criterion variables (and relevant predictor variables). Go through each hypothesis, present an inferential statistic and decide whether or not to reject the null hypothesis. Mention other interesting significant results. Include appropriate tables & charts.	**5**.After inferential statistics have been calculated.
Discussion	Explain your results. Speculate about causal relationships (with diagrams) and alternative explanations. Suggest how future research could resolve these questions.	**6**. After inferential statistics have been calculated. Keep adding references until the write-up is due, presented or submitted.
References	List all sources actually cited in the body of your write-up. Arrange alphabetically by authors' last names.	**3**. Start as soon as you have your first reference you intend to cite. Keep adding references until it is due.
Appendixes	Have this section if the guidelines require that tables, charts and diagrams be separated out from the body of the paper. For the Google Docs submitted in this course, put all of these in an above section.	**7**. After inferential statistics have been calculated.

Abstracts differ greatly in tone, length, and distribution, depending upon the audience. All abstracts should strive for brevity, clarity and objectivity. Journal articles want short abstracts so that they readers can decide if they want to read the entire article. Conferences want longer abstracts so that they can make a decision whether or not to include a presentation in a crowded program, and which section of other posters/papers a given presentation belongs in. The introduction section must be emphasized for interdisciplinary conferences (e.g., most student research conferences) because the people reading your abstract (even those making the decision about what proposals to include) are not content experts in your area. You have to let them know what you are talking about. Unless you are addressing such an audience (e.g., laypersons, scholars whose expertise is in other areas) that is unfamiliar with your topic, the abstract should not spend too much time on the introduction. You don't have to tell gerontologists "Alzheimer's disease is a widespread and devastating chronic brain syndrome" but that might be the opening sentence if the reader is not a mental health professional involved in elder care.

For most audiences, the major part of the abstract should be the *method* and *results* section of the write up: *how* did you do your study and *what* you found out. It should be clear from reading the abstract who was in your sample (sample size and important background characteristics), how the data were collected (laboratory, questionnaire, field count, archives), whether there was an experimental manipulation of a variable, and design (e.g., separate groups, repeated measures, sample vs. norms, correlational). If your statistical tests were unusual (e.g., not Pearson, t-test, ANOVA, Chi Squared) you might mention what they were.

The single most important thing to include in your abstract would be your major *findings* (results). Most abstracts contain no charts, graphs or tables, so you must be very clear with your use of words and numbers. Give us the descriptive statistics of the central tendency of the major variables, especially your criterion variable (i.e., percents, means, medians). Dispersion measures can also be included (e.g., standard deviations, ranges). Here are some example sentences.

“This sample had a median score of 23 on the Geriatric Depression Scale (range from 8 to 29).”

“Over half of the sample scored in the mildly depressed range, with 28 percent in the moderately/severely depressed range, and only 16 percent in the non-depressed range.”

Moderate and high correlations attaining statistical significance can be mentioned if they are important to your subsequent *discussion*. If the design was comparison (of separate groups, repeated measures or sample vs. norms) tell us the significant differences that tied into the major points of your discussion. (Make sure you run through the testing of hypotheses in your results section, but you don’t have to subsequently discuss all the hypotheses where you could not reject the null, and you certainly don’t have to discuss these in the abstract.)

Regarding the role of the discussion section in the abstract, unless you have an elaborate theoretical explanation (especially one that is innovative) you don’t have to spend more than a sentence on it in the abstract.

You don’t have to detail all the limitations of your research in the abstract (wait for the complete discussion section) and don’t bother with the obligatory “more research is needed” at this point (again, wait for the discussion section).

Since abstracts generally face the constraint of word limitations (which conference organizers strictly enforce) it is important to try to convey as much information in as few words as possible. It may be necessary to combine several sentences into one in order to reduce the total number of words.

Here is a video about writing abstracts.

Type of presentation	Project for this class	Article in academic journal	Conference (inter-disciplinary)	Conference (specialized)
Abstract length in words	**25 – 100**	25 - 200	100 - 1,000	100 – 1,000
Abstract emphasis	Method & Results	Method & Results	Introduction	Method & Results

Introduction. The next thing that the reader would see would be the introduction, which is frequently the longest part of the write-up. I suggest that you wait until you have your project (with hypotheses) actually approved, but you could start doing your literature review and use this section to take notes on the literature review as soon as you have committed to a particular topic. It is usually best to start with a major theory or historical figure in psychology: Piaget, Erikson, Milgram, Zimbardo, Mischel, Bandura, Festinger, Maslow, or Tversky & Kahneman. Review some of the major studies that have tried to apply these theories, but also the big question(s) that still remain unanswered. The purpose of this section is to show the relevance of your topic to the body of knowledge (theory + data) accumulated so far. Ideally, this will set up your hypotheses.

Another important thing to remember about your introduction is that it should be written as if you had written it before you performed your primary research. This is the way it usually is in a thesis or dissertation situation, where your advisory committee wants to see a thorough literature review and clearly formulated hypotheses before approving of the research methods you wish to employ. However, in practice, many researchers go ahead and collect the data before doing a complete literature review. (Even if you already have your results when you write your introduction, do not mention your results in your introduction. They come later in the section entitled *results*.) However much of a literature review you are able to do before you perform your research, realize that writing the introduction is a process of re-writing as you encounter more previously collected data and figure out how to creatively apply more theory.

If your class project is later accepted for presentation at a conference (or submitted for publication in a journal), you should definitely re-write the introduction to orient it more to the background and interests of that more specialized audience. For example, the poster "Autism vs. Asperger: did DSM-5 influence Google searches?" was presented at the Association for Psychological Science, so the introduction emphasized the symptomatology and nosology of those disorders. Had the poster been submitted to a conference about the internet, the introduction would have focused more on monitoring search engine activity.

Hypotheses. Next we have the hypotheses, clearly outlined so they can be easily set off for visual recognition. If this was a purely qualitative exploratory study (e.g., participant observation, focus group) then you can use some guiding questions that you started off with. Otherwise, you have to give at least one clearly stated hypothesis: a prediction of what one would expect to find (given the theories and previous data reviewed in the introduction). Usually, the number of hypotheses should be between two and five.

Look at a hypothesis as a promise of what your research will look into. So, once you have advanced a hypothesis, you are obligated to use your results to test that hypothesis.

It is tempting to immediately follow your statement of these hypotheses with a statement of something like “confirmed (p < .01)” but don’t jump the gun. The proper place to introduce your findings, and match them with your hypotheses, is in the results section.

This initial statement of hypotheses sets out the bare minimum of what your results and discussion will cover, but it does not prohibit you from going off into other directions (following an interesting, unexpected finding). So, your project may have begun with one hypothesis about dementia and depression and another hypothesis about dementia and paranoia. Suppose you could not reject the null for either of those hypotheses, but you found a strong correlation between depression and paranoia. You would report this in your results, and could focus your discussion on trying to find the most likely causal link between these syndromes, and suggest further research on this link.

Just remember that the hypotheses should be written before data are collected. It would be considered deceptive scientific writing if you conducted your research, used some data dredging statistics, found some interesting associations, and then developed the hypotheses at the end in order to pretend that you went looking for the findings that you stumbled upon.

Methods. Next comes a large section usually called methods. The purpose of this section is to describe, in excruciating detail, exactly what you did and how you did it. In general, it is best to error on the side of being too detailed. Only when your professor (or the journal editor) tells you to reduce the verbiage in this section, should you try to be more concise. Most authors prefer to divide this up into several subsections (and some journals require this).

The first subsection of your methods usually describes the who / when / where of your sample. This subsection might be called site, sample, subjects, or participants. You might have sentences such as these examples.

“Students (n = 50) at a community college in California’s Inland Empire comprised this sample of convenience. The participants were mostly female (62%), under age 25 (82%), never married (84%), and not yet parents (88%). A plurality was Hispanic (43%).”

“This archival investigation used anonymized records of nursing home patients (n = 83) in three proprietary nursing homes on the south side of Chicago. Participants were mostly female (76%) and had a median age of 83 (with a range from 62 to 101). Half were designated as African-American, and a quarter were Jewish. The remainder were mostly Roman Catholic (23%) of Irish, Italian or Polish extraction. Patients with diagnoses of dementia, or hearing or speech impairment were excluded from the sample.”

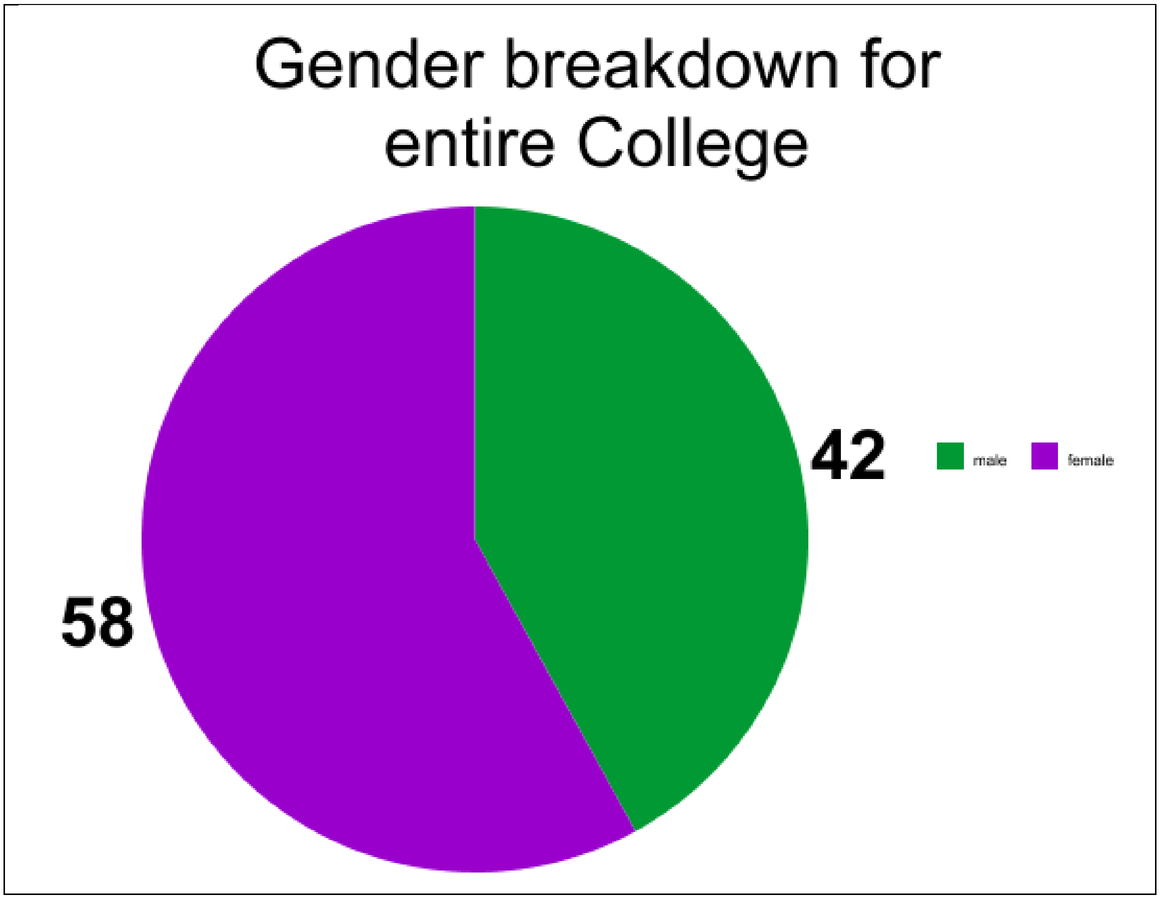
There is some disagreement about how many sample details should be reported here, in the methods section, versus the subsequent results section. My own judgment is that pure descriptions of background variables (e.g., gender, age, ethnicity, religious affiliation, socio-economic status) are appropriate in this section, but measures of dependent variables belong in the results section. The exception to the aforementioned rule would be if the measurement of the background variable discloses the hypothesis test. For example, if the hypothesis was that “Most of the participants in the skateboarding competition will be male” then don’t disclose the gender breakdown of your sample until you get to the next section, results.

If you do not have demographic measurements directly from your sample, you could use measures from the population from which they were drawn. Here is an example.

"Students (n = 50) at a community college in California's Inland Empire comprised this sample of convenience. The college's student population is mostly female (58%), under age 25 (69%), with a plurality being Hispanic (41%)."

To the extent that the sample was randomly selected and large, it could be expected to approach the norms of the population. Don't call your sample "random" unless you can describe the procedures by which you ensured randomization as described in this video. Otherwise, admit that you used a "sample of convenience" which means that you got your subjects by going to a place where you knew you could find cooperative people willing to participate.

Unlike the limitations of the abstract (i.e., words and numbers only), the main body of the write-up can include tables, charts, graphs and diagrams. This is especially true for the methods section (and later, the results section). Pie charts and bar graphs are particularly useful in describing the background variables of the sample (or the population from which it was drawn). Just make it clear whether it is the sample or population and what each slice stands for.



The next subsection of your methods usually describes the how you did the research. This section might be called *apparatus* / variables / data collection. The reader should come away knowing the details of how you operationalized each variable. If you had several variables, this requires several sentences (especially if the variables were measured on different scales). Here is an example.

“This one page questionnaire included measures of the aforementioned background variables (i.e., age, gender, ethnicity, parental status) as well as attitudes about the effectiveness of the job training (measured on a five level Likert scale) and job satisfaction (measured on a 0 to 10 ladder where 10 represented the ideal employment situation, and 0 represented the worst imaginable).”

Especially if you had a separate groups or repeated measures design, it is important that you clarify this in the apparatus section, and mention if and how any independent variables were manipulated.

“In this quasi-experiment, the sample was comprised of two pre-existing manufacturing departments, comparable in size (n = 24 workers), shift, geographical location (Midwest), gender breakdown (about two-thirds female), age (median around 35) and previous levels of training and productivity (as measured in units produced by individual worker per shift). The experimental group was assigned to human factors training, and the control group was not given such training until after this study was completed.”

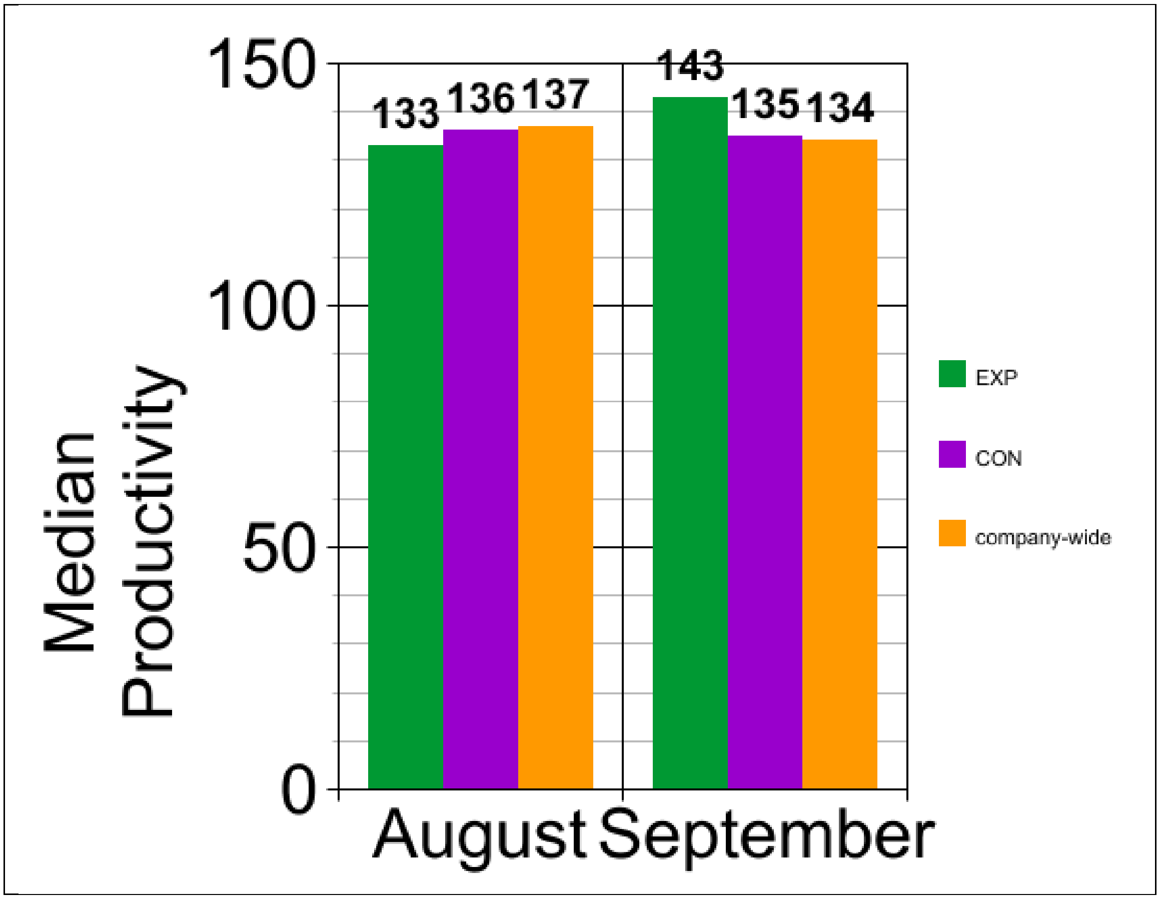
The last subsection of methods is sometimes called *procedure* or *data analysis*, and describes your statistical testing (but not the results of those tests, that is coming in your results section). Here you just tell how you coded the data and what statistical tests you did use (and why you chose those tests). Here are some examples of sentences you might use in this subsection.

“Of the fifty questionnaires initially distributed, all but one was returned. Two questionnaires were eliminated due to missing data on a criterion variable (subject attitude on training), yielding the current sample size (n = 47).”

“Given the obvious left skew in a criterion variable (productivity) and truncation on a predictor variable (age), normality was not assumed. The nonparametric measure of correlation employed was Spearman’s rho rank order coefficient. The nonparametric inferential statistic for differences between groups was the Mann-Whitney. Differences between this sample and company norms were tested with the Kolmogorov-Smirnov one sample test for maximum cumulative absolute difference.”

Results. Next comes the most important part of the write-up, the results (sometimes called findings). This should be the only place in your write-up where you say what you found out (except for a brief summary in the abstract). Begin this section by reporting on how the sample scored on the relevant predictor and criterion variable(s): central tendencies and dispersions (e.g., percents, means, medians, standard deviations, ranges). Use tables, graphs, and charts here (unless you are required to put them in an appendix), in addition to using words and numbers in the body of the write-up. Here is an example of such a table showing how the experimental (trained) group differed from the control (untrained) group in two months: August (before training) and September (after training).

	Experimental	**Control**	**Company-wide**
N	24	23	
Aug mean	129	131	132
Aug median	133	136	137
Aug percent meeting goal	56%	53%	57%
Aug percent exceeding company norm	42%	45%	50%
Sep mean	138	135	132
Sep median	143	135	134
Sep percent meeting goal	74%	52%	58%
Sep percent exceeding company norm	63%	48%	50%



After you have done this, go through each of the hypotheses initially advanced, and present relevant descriptive and inferential statistics. Be more detailed than you were in your abstract. Give the numerical scores for t, F or chi squared, Bayes Factors, as well as any degrees of freedom. You can also use tables, graphs and charts here. Here are four examples, one for each of the four designs.

Separate Groups

“H1: Workers receiving training will be more productive than workers who do not receive the training.

The experimental group (n = 24) had a median productivity of 143 units per shift, vs. the control group (n = 23) with a median of only 135 units per shift (Mann-Whitney z = 2.03, p < .05). Three-quarters (74%) of the experimental group met its production goals while barely half (52%) of the control group met those same goals (chi squared 4.21, df = 1, p < .01). Therefore, the null hypothesis can be rejected. These data are consistent with the explanation that the training was effective.”

Especially if the central tendencies (e.g., means, medians, percents) of the separate groups had not been given, then the effect size (e.g., Cohen d) should have also been listed.

Repeated Measures

“H2: Within the trained group, productivity will be higher after training.

None of the workers in the experimental group (n = 24) had a decrease in productivity following the month of training. Two had the same level of productivity and 22 (92%) improved their productivity (p < .001, according to the Sign Test for before and after). The median productivity score of 143 units after training, was significantly higher than the 133 units per shift measured before training (p < .01, Wilcoxon Rank Sums Test). Therefore, the null hypothesis can be rejected. These data are consistent with the explanation that the training was effective.

By contrast the control group (n = 23) saw a productivity increase in only 10 workers, a decrease in an equal amount and three unchanged. The before training median productivity of 136 did not differ significantly from the after training median of 135 (p > .10, Wilcoxon Rank Sums Test). Therefore, the null hypothesis cannot be rejected for explaining repeated measures within the control group."

Usually, effect sizes are not used with repeated measures designs, but what is becoming more popular is a 95% confidence interval for the difference between the means.

Sample vs. Norms

"H3: The productivity of the trained group will be higher than company norms for comparable departments.

The company mean productivity was 132 (S.D. = 3.4) before training, while that of the experimental group was close at 129 (S.D. = 2.5). This was not a significant different (t = 1.05, df = 23, p > .10). The company median was 137, and only 42% of the experimental group exceeded that median (p > .10, Sign Test).

However, when we look at the mean performance of the experimental group after the training (138, SD = 2.6) it is higher than the company-wide mean (132, SD = 3.2, t = 2.79, df = 23, p < .05). Now, 63% of the experimental group exceed the company-wide median of 134 (p < .05, Sign Test). Therefore, the null hypothesis can be rejected."

Correlational

"H4: The same workers who were the most productive workers before training will be the most productive workers after the training.

Within the experimental group (n = 24), those who scored high on August's productivity also scored high on September's productivity (rho = +.64, p < .01). The null hypothesis may be rejected.

Of the workers meeting their production goal in August before training, all of those met their production goal in September after training. All those who failed to meet production goals after training, had failed the previous month (chi squared = 11.6, df = 1, p < .001). The null hypothesis may be rejected. "

For correlational designs, it is also possible to report the 95% confidence interval on the strength of the association, such as "+.34 to +78."

After you have gone through each of your initial hypotheses, you may report on additional findings (i.e., things that you discovered but you were not initially predicting). Perhaps in the above example you found that it was the female workers (more so than their male counterparts) who really benefited from the training. You might use the following words, tables and charts.

"Within the experimental group, it was the women who scored the highest productivity gains.

	N	Sep Median Productivity	Percent meeting goals	Percent exceeding company norms
Males	8	138*	60%*	50%
Females	12	148*	80%*	67%

Their mean productivity post-training was higher than that achieved by the men (Mann-Whitney z = 1.98, p < .05) as was the percent meeting production goals (chi squared 4.04, df = 1, p < .05) but the difference was not significant for the percent exceeding corporate norms (chi squared 2.18, df = 1, p > .10)."

Discussion. The last major section of the write-up is the discussion in which the purpose is to figure out what all the findings mean (and where should we go in the next phase of research). Do not just repeat your results; now you must explain them. You already said that the experimental group did better (better than the control group, better than company norms, better than the same group did in the previous month), now tell us why. What was it about the training that worked? You already told us that the trained women did better than the trained men? How could that be?

Don't be afraid to speculate beyond the data that you have. You can tie in other sources of previously accumulated data. You can apply other theories not covered in the introduction. Use diagrams to show different causal relationships between correlated variables. Speculate about the role of *moderating* variables: *mediation*, *potentiation* and *attenuation*. Unless you performed an experiment in which you manipulated variable X and observed a difference in variable Y, be cautious about claiming proof that X causes Y. Always consider other plausible relationships. Could Y cause X? Could the correlation be *spurious*, with both X and Y being caused by another, unobserved factor Z. Remember, just because a variable is independent in terms of it being a background factor for the subject, or something that happened to the subject independently of his/her preference, it may still be due to some other background factor that also produces the dependent variable.

For example, for almost a hundred years, many pediatricians and obstetricians have advocated routine infant circumcision of baby boys. One reason has been epidemiological data that sexually transmitted infections were lower among circumcised men. This led to the assumption that circumcision toughens the penis and makes it more resistant to such infections. But these epidemiological data were not the result of some randomized experiment in which half of European infants were forcibly circumcised and the other half served as a control that would be denied circumcision. Whether or not the infant was circumcised was determined by the preferences of the parents and perhaps the protocols of the hospital in which he was born. Jewish parents and hospitals circumcised the boys while most Gentile males managed to escape the process. So, was the lower incidence of STIs due to some effect of the circumcision on the penis, or was the lower STI rate due to the fact that Jewish men were less likely to have sex with infected women? Perhaps Jews were more monogamous. Perhaps Jews were less likely to visit prostitutes. Perhaps Jewish women were less likely to carry infections. Perhaps Jewish men used condoms more consistently. We must be careful to avoid the *post hoc* fallacy in the interpretation on non-experimental data.

One of the major decision points in a write up is where to put each citation of previous research. Does it belong more in the introduction or the discussion? Frequently, I see students initially putting some reference in one section, and then cutting and pasting to the other. This is fine. The deciding factor should be "does the citation serve more to set up the hypothesis or explain the results"? In general, where you have significant findings (i.e., found confirmation for your hypotheses), it is better to backload and use these sources in your discussion to explain your findings. When the opposite is true, and you were unable to reject the null, it is best to frontload so that your hypotheses at least seemed plausible ones to begin with.

At the end of the discussion section comes a couple of obligatory paragraphs (which may be as short as a sentence or two each). One deals with the limitations of your study. This is the question of lack of *internal validity* and *external validity*. The former goes back to something that should have been covered in the methods section: problems with the operational definition of the variables and/or limitations in the design. Here are some examples coming from a study of the impact of bonuses on the performance of salespersons.

- The criterion variable, effectiveness of the salespersons, was subjectively assessed by their supervisors, and there was no previous validation procedure, or any opportunity to study the inter-rater reliability of that measure.

- The predictor variable (time as a salesperson) lacked normality, suffering from a floor effect leading to left truncation and a right skew.

- The use of a quasi-experimental design opened up several potentially confounding variables, especially since the different sales teams were headed by different supervisors who might have rated their teams according to different standards.

- The experimental manipulation (an incentive of a set of steak knives) given to one of the groups may have been inadequate to motivate higher performance.

- The time period between the repeated measures (over 13 months) may have been too long, such that whatever immediate impact the incentives had, that had faded away by the time the follow up measurement of performance took place the following year.

The last part of the discussion section is suggestions for future research. You need to go beyond the saying “more research is needed” (which is the equivalent to repeating the truism “science is never finished”). You need to say what type of research should be done. Suggest specific modifications of your study (e.g., new measurements of key variables, different populations). One thing which general fits is that if your research was only a survey, suggest how a true experiment would be able to resolve some of these questions in the future. If your results were significant, you could also suggest that the next cycle of research be qualitative (e.g., focus groups) which could examine some of the dynamics of the participants' decision making.

Citations & References

The last part of the write-up would be the list of references. In this class, and for many conferences and scholarly journals, the format for the references and their citation within the body of the write-up, is known as *APA style*, because it is based on the *Publication Manual of the American Psychological Association*. Complete information can be found here at this site, but here is a quick and helpful guide from Purdue's Online Writing lab.

There are many rules for APA format, and you will only master most of them over time and with much practice. The most important thing to start with is how to list references and how to cite those references. Begin by building your list of references (which is placed at the end of your write-up), as soon as you begin your literature review.

This list is called *References* (not bibliography) and is to be organized alphabetically by authors' last names. We do not organize by the chronological order (when the citation appears in the main body of the write-up). We do not organize by type of source: we don't put all the journal articles together, then all the books, then all the conference presentations. We do not organize by title of the article, or by name of the journal. Authors' last names is the way to organized all sources.

If we have more than one source from a particular (first) author, (say four things: a couple of articles, a conference presentation, a chapter in a book) here's how we decide which one should come first. All those sources authored solely by that individual come first, and then we include the co-authored pieces, organizing them alphabetically by (the first) co-author's last name, then by the next co-author's last names, etc.

Where there are further ties, we organize by year of publication (oldest first). When we have two publications by the exact same set of authors coming in the same year, we organize by alphabetizing the title of the article, book, chapter, or presentation. Then we rename the date of publication with a small a, b, or c afterward. Here are some examples.

Brink, T.L. (1978) Geriatric rigidity and its psychotherapeutic implications. *Journal of the American Geriatrics Society*, *28*, 274-277.

Brink, T.L. (1979) *Geriatric Psychotherapy*. New York: Human Sciences Press.

Brink, T.L. (1999a) Case study method. In D.G. Benner and P.H. Hill (Eds.) *Baker Encyclopedia of Psychology and Counseling* (2nd ed.) p. 173, Grand Rapids, MI: Baker Book House.

Brink, T.L. (1999b) Midlife crisis. In D.G. Benner and P.H. Hill (Eds.) *Baker Encyclopedia of Psychology and Counseling* (2nd ed.) pp. 752-754, Grand Rapids, MI: Baker Book House.

Brink, T.L. (1999c) Qualitative research methods. In D.G. Benner and P.H. Hill (Eds.) *Baker Encyclopedia of Psychology and Counseling* (2nd ed.) pp. 997-998, Grand Rapids, MI: Baker Book House.

Brink, T.L., Yesavage, J.A., Lum, O., Heersema, P.A., Adey, M., Rose, T.L. (1982) Screening tests for geriatric depression. *Clinical Gerontologist, 1,* 37-43.

You must include complete bibliographical information for each source. You are going to integrate together all of your sources, whether they are from conference presentations (e.g., posters or oral presentations), periodicals (e.g., scholarly journals, magazines, newspapers), websites, books, chapters in books or encyclopedias. When in doubt about what to include, include more information rather than less.

For articles, include in this order: name(s) of author(s), date published, title of article, name of journal, volume, number, page numbers. If available, you may include the digital object identifier, which is the site where the article can be found on the internet. Everything should look like this 1982 reference.

Brink, T.L., Yesavage, J.A., Lum, O., Heersema, P.A., Adey, M., Rose, T.L. (1982) Screening tests for geriatric depression. *Clinical Gerontologist, 1*, 37-43.

Books cited should include, in this order: name(s) of author(s), date published, title of book, edition (if there are more than one), city of publication, name of publisher. Everything should look like this 2013 reference.

Carmody, D.L. and Brink, T.L. (2013) *Ways to the Center: an introduction to world religions.* (7th ed.) Belmont, CA: Cengage

Within an edited book where individual chapters have separate authors (or in an encyclopedia), include in this order: name(s) of author(s), date published, title of article or chapter, editors, title of book, edition (if more than one), city of publication, name of publisher, volume number, page numbers. It should look like this 1999 reference.

Brink, T.L. (1999c) Qualitative research methods. In D.G. Benner and P.H. Hill (Eds.) *Baker Encyclopedia of Psychology and Counseling* (2nd ed.) pp. 997-998, Grand Rapids, MI: Baker Book House.

Conference presentations, whether oral or poster, should also be referenced, including names of authors, date, title of presentation, name of sponsoring organization, location. It should look like this 2007 reference.

Waters, N.A. & Brink, T.L. (2007) Secularism: development of a scale. Sociedad Interamericana de Psicologia, Mexico City.

Because you are giving complete bibliographical information in your reference list at the end, there is no need to give all that information right in the body of your write-up.

Every source within the reference list at the end must be cited, at least once, in the body of the report (e.g., in the introduction or discussion). Within the body, only cite by authors' last names and the date of publication, not by academic institution. The rest of the bibliographical information (e.g., title, journal, volume, page) belongs in the references, not in the body of the paper.

Not like this: "A few years ago, John W. Jones, M.D., Ph.D., F.R.C.S., distinguished professor of psychiatry at Johns Hopkins University, proved our point in an article entitled Baby Blues, published in the *American Journal of Psychiatry*. It was his opinion, that most mothers get basically depressed sooner or later."

Like this: "Jones (2013) found that 54% of mothers report at least one major symptom of depression within a month after having given birth."

Readers can get the complete information about the article by looking at the reference list.

Other things to remember about this formal APA style …

- use many references, not just a couple
- avoid long quotations, just paraphrase (but still cite)
- when referring to authors by name, outside of the parentheses, include the word “and” between co-authors, but when their names go within the parentheses, used the & symbol
- when there are more than two authors, after their first citation, just use the first author's name followed by the abbreviation *et al.*

Don’t confuse these words: *sight* is something to see; *site* is a location, perhaps on the web, but *cite* is a verb meaning to acknowledge a reference, and citation is the noun.

Oral Presentations

Another way of presenting the report would be orally, usually at a professional conference. Start with the guidelines furnished by the organization at which you will present, especially when it comes to the amount of time you have to present. It is alright to take less time, but do not go over whatever limits you have been given.

There are two mistakes novice oral presenters make. One is trying to read everything to the audience. If you time it out before hand, and read fast, this may keep you on the time schedule, but it is incredibly boring. The other extreme is also a mistake, using no notes and just trying to speak off-the-cuff. You will come off as unprepared and will probably go over the time limit without covering everything because you will get side-tracked by irrelevant details.

The happy medium is to use a slide presentation format (e.g., Power Point, Prezi, Google Slides) to keep you on track. Include many pictures, charts and diagrams to keep the audience interested. Make sure that the type font is readable from the back of the room. Have your first slides be the title, your name, affiliation, and contact information. Have your last slides repeat this and have Quick Response (*QR*) codes to your data in Google sheets so that the audience can view (but not edit) a spreadsheet with your data and also a document with your complete write-up (and especially the references which are usually not completely included in the oral presentation).

On the day of the conference on which your presentation is scheduled, arrive in the room 10 minutes early. Identify the "chair" (the person who has been selected by the program committee to lead this section). Clarify these points.

- Have the computer and projector already been set up? If not, get on that immediately.

- Have speakers for each of the scheduled presentations arrived? (If someone is a no show, will the chair read that presentation or does that mean that the rest of the presenters get extra time?).

- How many minutes will each presentation get?

- How will the chair signal the time? (time remaining works best).

- Should questions and comments from the audience be taken after each presentation or held to the end? My preference, both as a chair and as a speaker, is to hold these to the end. The quality of the resulting discussion is much better because we can see the thematic links between the different presentations. Another problem with taking questions / comments after each presentation is that too much time is taken up by one of the early ones, and the person who has to give the large presentation is rushed to finish before the next session starts.

In order to avoid technical problems with your presentation, do the following (in order of importance).

- Save your presentation as a Power Point file on a USB in case there is no internet in the presentation room. Once you have identified which computer will be used in the presentation, upload your Power Point file to the desktop of that computer and then put the USB drive back in your pocket. (At every large conference, dozens of such drives are lost by presenters who leave the room when the session is over, forgetting to get their USB drives out of the computers.)

- Before you leave for the conference, upload your power point file as Google Sheets. You might lose the USB drive (especially at the airport security checkpoint). This is also a helpful backup if you end up with a computer at the conference that cannot accept or read your USB drive, or one who lacks Power Point.

- Convert your Power Point file to a pdf file. Many computers lack Power Point, but most can view a pdf.

- Upload the Power Point file to your own laptop and bring it to the conference in case there is no computer available in the room of the presentation.

- Bring a connecting cable, extension cord, power strip, and charger cord with you to the presentation. Sometimes these are not available and your laptop does not have a full battery because you have been using it more than usual. I have even known presenters to bring their own projectors and screens just to make sure an audio-visual presentation can be made.

When it comes time for questions and comments from the audience, relax. There are two kinds of people who respond to an oral presentation. The vast majority would be those who are informed, sincere and helpful. They will give good advice and constructive criticism so that your next cycle of research can be better. There are a few in the audience who are old curmudgeons or brash students wanting to show off how much they know (and a little jealous that they are not up there making a presentation). Whichever kind of response you get, remember that the audience will not remember what they said, or even what you said, but how you looked when you responded.

So, here is the proper way to deal with audience comments to your oral presentation. Look directly at the person who is asking the question or making the comment. Pretend you are a Rogerian therapist, maintain eye contract and nod your head every ten seconds or so. When it is your turn to respond, try some of these.

- Sincerely thank him (it is usually a male, especially the difficult ones) for his comments.

- If possible, try to present some additional information about your data or details about the design, or perhaps some additional findings in your primary or secondary research on this topic. Use this phrase to lead in. "I can see why that question arose, because we really did not have time to report on several things that would have clarified these gaps. I better take this opportunity to tell you more about our research."

- Reverse the roles by asking a question of the person making the comment. "What do you recommend that we do about …" If there is a critique of the design or choice of statistics, ask him to clarify a good alternative approach. (This is where some questioners will go off talking about the virtues of Structural Equations, Bayes Factors or nonparametrics.)

- Praise him for his superior knowledge in this area and offer to discuss this after the session so that he can give you more guidance.

- Invite the questioner to view your data file or to comment on your complete write-up by using the Quick Response Codes. The audience will not remember what you said, but may remember the poise with which you said it.

If the questioner is really an expert with good ideas, you will learn from this encounter. If the questioner is a "blow hard" the rest of the audience will figure it out and you will look great handling him so kindly.

Poster

Another form of conference presentation is the poster. This is similar to a "science fair" presentation you may have done in k-12.

As soon as you know that you will doing a poster presentation, check the organization's guidelines for size and type. Will they provide easels, cork boards, or hard surface panels? If it is easels, you will need to mount on a portable board. If it is a hard surface, bring a roll of masking tape.

If it is corkboard, bring push pins. Just in case, pack both push pins and masking tape. You will need at least two (preferably four) pins per standard page, and a pin every foot for larger posters. So, a three foot by four foot poster may require 14 pins because you will start at one edge with four pins, then unroll it about a foot, one pin on the top and one on the bottom, unroll another foot, another pin on the top and one on the bottom, then also for another foot, and at the end put four on the other side of the poster. Always bring extra pins, because some will get lost, and you can make friends by sharing pins with presenters who need some.

Consider how you will be traveling to the conference. If you are flying to the conference, call the airline before you make your reservation and ask if it is alright to bring a rolled up poster (or folded board) in the overhead compartment. (Usually, Southwest is OK with this, but other airlines can be problematic charging you extra for a checked piece of luggage.) If you put it up in an overhead bin, give yourself some reminder so that you do not leave the plane without it.

Here's how you write the poster. You will have a shortened version of each of these sections: abstract, introduction, hypotheses, method, results, discussion, references. You will need pictures (either photographs or images from the internet), one for the school logo and several related to your topic. You definitely need tables, charts, graphs and/or diagrams.

Use different font sizes, perhaps 96 for Title, 72 for your name, your advisor's name (unless he/she is listed as a co-author), your institutional affiliation (and logo), your email address, 60 for each of the aforementioned sections, and then 36 for the actual words (only the tables and references should be much smaller than this).

Put Quick Response Codes on the bottom: one for a link to a Google Spreadsheet (view only) that contains your data, the other for a link to a Google Document file of your write up (which allows anyone to view & comment).

The organization of your poster will be similar to that of the written report or slide show presentation. At the top should be the title (in the largest font), then your name(s), affiliation (and perhaps its logo).

If you want a high quality professional looking poster, it will cost about a hundred dollars at Kinko's, more if you want it laminated. The cheapest way to do your presentation (and the easiest way to carry it) would be a dozen 8.5 by 11 inch sheets, then taped or tacked to a presentation board.

Start out writing the poster in word, but if you are going to get a commercial poster printed, there is a special Power Point template to use: one big slide.

When you arrive at the conference, attend one of the poster sessions prior to yours in order to see how it works. If there are assigned locations (e.g., board #47) locate where yours will be. Arrive at your location fifteen minutes before your session starts. Someone from a previous session might still be at that board. When the board is open, mount your poster. If your poster contains numerous small sheets of paper, start putting up the one with the title and your affiliation and have the last one be the references and QR codes. If it is one piece of paper or vinyl rolled up, start over at one side and put in four pins all the way down and then unroll it, putting in more pins top and bottom as you unroll.

When the poster is up, stand to one side, smiling at passers-by. Expect that the majority will simply walk by. That is because they only have so much time to see all the posters and are looking for the ones that best match their interests.

Be prepared to give a two-minute walk through if someone requests. Don't read it to them, but make it a guided tour of the poster. Point out and explain the pictures, tables and charts (without using the word "basically" or phrases like “in my opinion”). Clearly explain the operational definitions of your key variables. For both the written poster and the oral summary/answers, try to avoid informal language, pejoratives, and euphemisms (e.g., "issues").

Some people will have comments, suggestions or questions. Do not be offended by anything they say. Smile and thank all those who do have something to say. Invite them to view your data file or to comment on your complete write-up by using the Quick Response Codes.

Here is a short video of a presentation of Crafton faculty at a poster session of the Association for Psychological Science (APS) in Chicago.

The Proper Tone

The tone of scientific reports is usually formal, respectful, and restrained. You are not trying to impress, much less to insult, but to inform (and possibly inspire). That means avoid slang, euphemisms and pejoratives.

AVOID	EXAMPLE	USE THIS INSTEAD
Euphemisms	Special needs	Use the specific diagnosis, such as autism
Pejoratives (or any term that could be perceived as such)	Neurotic	High on neuroticism
Sexist language	Fireman	Fire fighter
Insulting other authors	"Jones and Smith are idiots, their data are incredible, their theory is ludicrous and their conclusion is asinine."	"One question about data validity would be … " "An alternative theory could be that ..."
Bragging & exaggeration	"Not only does our study disprove all the previous data, but our theory revolutionizes this field."	"The findings are limited to … " "Future research could further substantiate these findings by … "
Inappropriate plurals	"This data was … "	"These data were ..."
Unnecessary terms	Basically, in our opinion, we felt that	We suggest, we conclude

The final arbiter on guidelines for length, format and tone is always the audience or readership to which your report will be presented. If it is an article for a scholarly journal, obey the editor. If it is a presentation at a conference, obey the guidelines of the conference organizing committee. If it is a project for a class, follow the instructions of the professor.

One last thought, when you are at a conference presenting your research, never look at anyone else as an enemy. No one is out to get you. No one is going to steal your research. Your goal should not be to hide it from public view, but to disseminate your work as widely as possible. Everyone you meet at a conference is a potential ally: someone who could publish on a similar topic and cite your work, and even be a colleague on a future research project.