READING SOCIAL SCIENCE METHODS

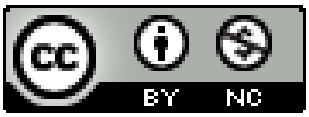
ANN REISNER

Reading Social Science Methods

Reading Social Science Methods

By Ann Reisner

WINDSOR & DOWNS PRESS
URBANA, IL



Published by Windsor & Downs Press, Urbana, Ill., part of the Illinois Open Publishing Network (IOPN). IOPN is the press of the University Library, University of Illinois at Urbana-Champaign.

Published as part of the OPN Textbook Series.

ISBN (Online): 978-1-946011-19-0
ISBN (PDF): 978-1-946011-20-6

Please cite this book using the DOI: https://doi.org/10.21900/wd.18.

This book was produced with Pressbooks (https://pressbooks.com) and the PDF is rendered with Prince.

Special Permissions

The following are used by permission, and all rights are reserved by the copyright owners:

Figure 1.3, "Schoolboys," is used courtesy of the National Library of Ireland.
Figure 1.5 reprinted from Erin E. Buckels, Paul D. Trapnell, and Delroy L. Paulhus, "Trolls Just Want to Have Fun," *Personality and Individual Differences*, The Dark Triad of Personality, 67 (September 2014): 97–102, with permission of Elsevier.
Table 8.1 reprinted from Roberta Weiner et al., "Climate Change Coverage in the United States Media during the 2017 Hurricane Season: Implications for Climate Change Communication," Climatic Change 164, no. 3-4 (February 2021), https://doi.org/10.1007/s10584-021-03032-0, reproduced with permission from SNCNC.
Figure 12.1 used in the web edition only courtesy of Moviestore Collection Ltd/Alamy Limited.

Cover design by Elizabeth Budd. Cover image derived from Christopher (shutterhacks), licensed CC BY 2.0.

Funding Statement

Work on this textbook was partially funded by the 2021 Faculty OER Incentive Program, funded by the University Library, Office of the Provost, and University of Illinois Student Government.

Contents

PART I

GETTING STARTED

1. Introduction

Science (including its practical older sibling, engineering) has been a driving force for civilizations for centuries—arguably since humans learned to control fire and absolutely since the Age of Enlightenment. Science and engineering have altered virtually all aspects of what we are; what we wear, how we eat, how we protect ourselves, how we get from one place to another, how long we live, and how much pain we live in. **Most of these changes are beneficial.** Smartphones are great, binging on Netflix (or whatever) is a fine way to spend a lazy night, and having all the knowledge of the internet is fantastic. But the benefits come with risks.

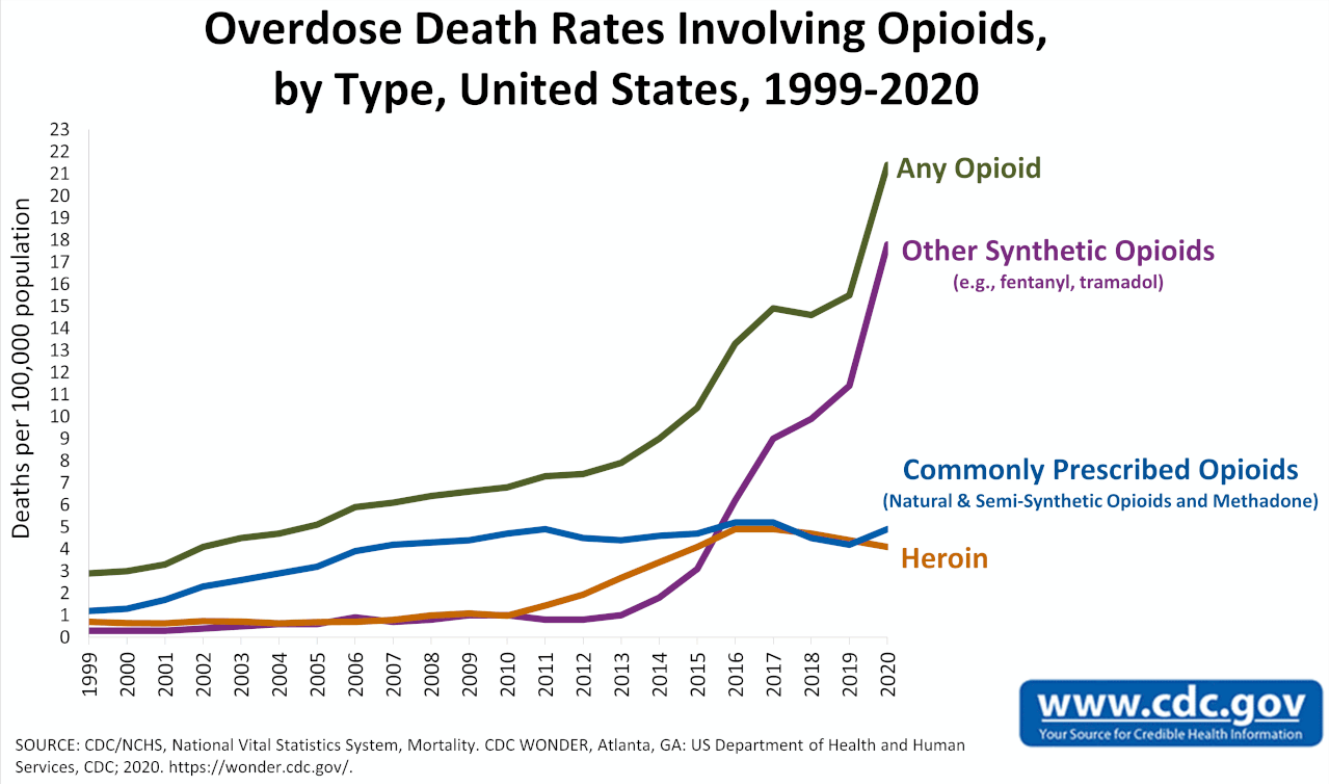


Figure 1.1. Overdose Death Rates Involving Opioids, by Type, United States, 1999-2020. Center for Disease Control and Prevention, 2020. Public domain.

Percocet, a combination oxycodone/paracetamol, is an opioid/non-opioid pain reliever used for moderate to severe pain.[1] When you have severe pain, Percocet is a blessing. It is also dangerous. Percocet can cause fatal complications even under a doctor's supervision. Taken for pleasure outside of the medical system, Percocet and other opioids kill (42,000 lives in 2016 alone; see Figure 1.1).

In much the same way, my iPhone[2] is a wonderful thing. I use my phone for notes, music, games, light, texting, directions, internet, camera, fitness timers, and—rarely—a phone. Smartphones are so integrated into daily lives that businesses frequently assume that their customers will have one. Essentially, they require that assumed phone to verify purchases (credit card checks), identity (internet security checks), emergency messages (schools and hospitals), or that the cable guy or the plumber is actually on the way.

Nevertheless, just because smartphones are great does not mean that they are always great. Texting while driving increases car accidents. Texting while in lecture decreases grades. For example, what would you do if research suggests that having the phone out while talking to friends decrease the satisfaction of interacting with friends?

Shalini Misra and her colleagues looked at how mobile phones affect the quality of social interactions in public.[3] Because the researchers were interested in real-world conversations, they asked groups of two who were waiting to order drinks in coffee shops if they would be willing to participate in a research study. If they agreed, they were seated on two chairs with a table between them. Half of the dyads in the study were asked to talk about "their thoughts and feelings" about plastic Christmas trees; the other half were asked to talk about the most meaningful events of the past year. A trained observer sat relatively far away and watched whether or not the participants had phones out during their conversation.

After ten minutes, both participants were asked to rate their satisfaction with the conversation. The mere ***presence*** of a phone dragged down the quality of the interaction. When one person in the conversation pulled out a smartphone, the participants felt that they were less connected to their partner and that the conversation was less fulfilling compared to conversations with no phone. The appearance of the phone hurt more for close friends than casual acquaintances. Satisfaction went down whether the participants were discussing plastic trees or the most important event of their lives.

1. "Percocet—oxycodone hydrocholoride and acetaminophen tablet" (Archived Version), DailyMed, National Library of Medicine, National Institute of Health, last modified July 2010,
2. This is not an endorsement for the iPhone. It's just what I happen to have, and it works for me.
3. Shalini Misra et al., "The iPhone Effect: The Quality of In-Person Social Interactions in the Presence of Mobile Devices," *Environment and Behavior* 48, no. 2 (February 1, 2016): 275–98, https://doi.org/10.1177/0013916514539755.

Does this study mean that you should toss your phone out? No, not really. Does the study suggest that you should put the phone away when you are talking with your friends? No. "Should" is a verb signifying a moral directive. If you do not care about how your friends feel, then why should you put your phone away? Do what you want.

If you do care about the quality of your interactions with close friends *and* you think that the study is sound, then—yes—Put. The. Phone. Away.

Again, you need to put the phone away *if* you think that the study is sound; that is, if you accept that the researcher appropriately followed the rules of gathering evidence. This book is based on the belief that knowing how to judge the validity of scientific papers and using science as important background for making decisions are really important and useful skills to have. When you understand the rules of science, you have developed the context to decide for yourself when, how, and *why* to control your media use so that your media does not control you.

What Is Science?

Science is a set of rules on how to find things out.[4] These rules vary with every different scientific procedure, but every method is defined by a specific set of rules. Simplistically put, any scientific method is valid to the degree that the researcher follows those rules and weak to the degree that it does not. Therefore, the first task of a skilled reader is to learn what those rules are.

The rules tell you such things as:

What you can and cannot study using a particular scientific method or instrument. In medicine, doctors do not use a thermometer to measure blood pressure and do not give blood transfusions to cure a cold. Doctors select the tool that is most useful for what they want to measure. The principle is the same for social science, and both researchers and readers should know what each social science method is useful for—what the method can study and what it cannot.

What population the research can, and cannot, make claims about. The purpose of social science is to make claims about some group of people or things (for example, the population of fruits in Figure 1.2). But researchers can only make claims about the population they actually studied. If a physical scientist found out that fruits help people decrease stress, then an intelligent—and stressed—reader might want to start eating more fruits. So far, it is all very clear.

4. —and the findings that come from applying these rules.

However, if you started looking at exactly what the scientist studied and—in this example—he only looked at bananas, then all you can really take from the study is that eating bananas will reduce stress. The researcher extended his results too far when he said "fruits." Since the researcher did not look at apples, he (and you) should *not* assume that apples reduce stress. They may; they may not. You do not know. Of course, the difference between bananas, apples, and even pears is reasonably straightforward and hard to confuse.

Figure 1.2. Fruits. Photo by Rodrigo dos Reis, Unsplash, Unsplash License.

The social science version of this mistake is a bit trickier, particularly with humans' self-centered tendency to forget that other groups of people are not exactly like them.

Figure 1.3. Schoolboys. Photo by A. H. Poole Studio. Poole Photographic Collection, National Library of Ireland.

For example, let us say that a researcher is looking at whether reading graphic novels changes teenage behaviors. In her conclusions, she recommends *policies that high schools should adopt* to prevent harmful effects from graphic novels. Following the researcher's recommendations is reasonable, assuming that you trust that the researcher studied students who are similar to the students the policy is being developed for; however, what if the sample studied were the students pictured above? (See Figure 1.3.)[5] Can she reasonably say that these teens fairly represent all teens? Are some teenagers left out—African-Americans, Asians, females, or white males who wear tight jeans to school? Would you want to be subject to a policy made for white boys who go to private school (particularly schools that encourage wearing shorts and a blazer with white piping)?[6] Policies developed on the reactions of privileged, white boys might not be suitable for other groups.

5. The photograph is not of the actual study population.
6. In this culture, shorts and blazers with white piping signify upper-class and wealthy attire for elite private schools, a very narrow class of people in the United States.

Why Care about These Rules?

Essentially, using science means agreeing to accept a set of findings to the same degree that you trust the methodology. This means several things, all of which are important. First, you can (and should) decide on whether you are going to accept the findings *before* you know what the findings are.

If the methodology is crap (no matter how much you like the findings), then the findings are crap. If the method is good, then you ***must*** accept the findings as true, even if you hate what they imply.

Example: The Dark Tetrad

Let us say that you really like trolling online—the snarkier the comment is, the better.

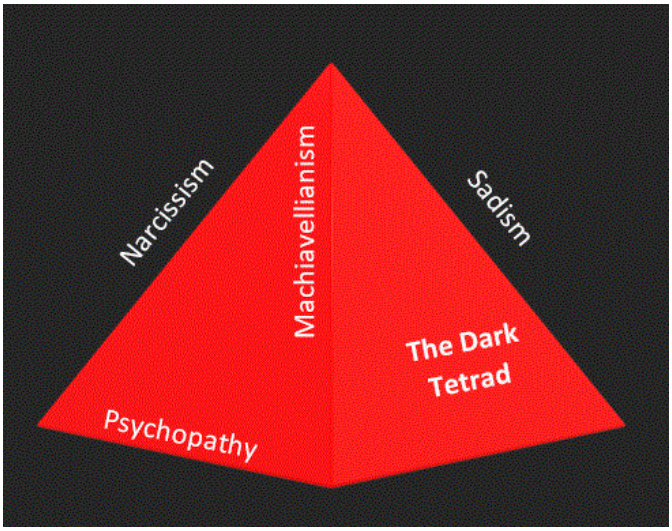


Figure 1.4. The Dark Tetrad.

At the same time, you think that you are a nice person, an asset to your community, and a faithful follower of your religious principles. You are committed to both trolling and your self-image.

Buckels, Trapnell, and Paulhus' study of internet trolling challenges the idea that trolling is a harmless hobby.[7] In a series of studies, Buckels and his colleagues looked at whether internet trolls show the personality characteristics of the Dark Tetrad, a series of highly negative personality characteristics: Machiavellianism (the willingness to manipulate and deceive others), narcissism (egotism and self-obsession), psychopathy (the inability to show remorse or empathize with others) and sadism (feeling pleasure in the suffering of others) (see Figure 1.4).

The study found that trolls were far, *far*, ***far*** more likely to be card-carrying members of the Dark Tetrad club than any of the other tested groups.

7. Erin E. Buckels, Paul D. Trapnell, and Delroy L. Paulhus, "Trolls Just Want to Have Fun," *Personality and Individual Differences*, The Dark Triad of Personality, 67 (September 2014): 97–102, https://doi.org/10.1016/j.paid.2014.01.016.

Now you have a conflict—you think that you are a nice person, but you really like to troll (and, to be fair, you have to admit that you enjoy getting a rise out of people). Figure 1.5 shows that trolls are not nice (for an important qualification, see footnote[8]), so one of these two things has to be wrong. Either you are not a troll, or you are not as nice as you think you are. In this case, science has challenged you to look beyond your biases about yourself and consider some hard truths, even if you hate it.

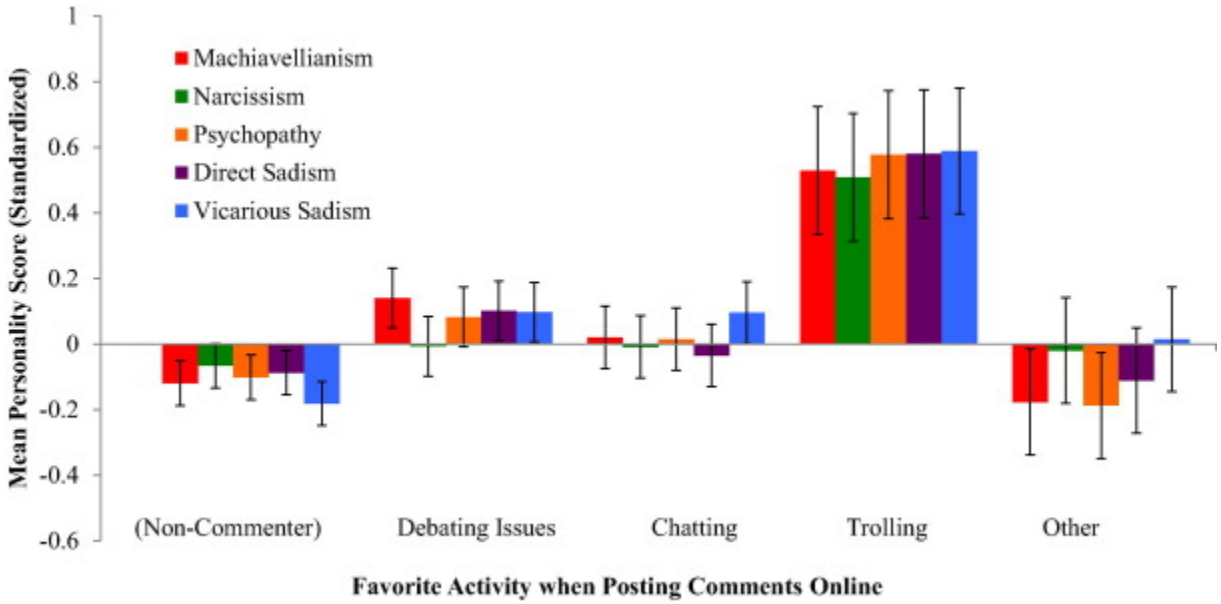


Figure 1.5. Dark Tetrad scores as a function of favorite activity when posting comments online. Reprinted from Erin E. Buckels, Paul D. Trapnell, and Delroy L. Paulhus, "Trolls Just Want to Have Fun," Personality and Individual Differences, The Dark Triad of Personality, 67 (September 2014): 97–102, with permission of Elsevier.

As illustrated above, science is—or should be—a way for you to break through your own biases. We humans have considerable psychological defenses, and we use them to defend our own prejudices, including the idea that we are personally exceptional. Communication science has numerous examples of situations in which people think they are much less likely to be influenced by media than they actually are:

8. The researcher used a survey, a method that *cannot* prove causality. A survey can only suggest that trolling and Dark Tetrad personalities appear together (covariance).

Table 1.1. Examples of Cognitive Biases in Interpreting Impacts from Media

Examples of Cognitive Biases in Interpreting Impacts from Media	
Example of situations where people do not think that they are affected, but research show they more likely are.	**Theory explaining the distortion**
Other people are affected media messages, but I am not.	Third person effect
Media images do not affect me.	Priming
Reading stories of a robbery is not going to affect how much I blame the victim for being robbed.	Effects of reading episodic news

Science Is Also a Way to Get Deeper into a Conversation

Science is, or can be, a way to move beyond starting and ending arguments by trading personal opinions: the "Yes, you are"/"No, I'm not" quarrel.

Let us take a concrete example. Is the picture below sexist?

Figure 1.6. Marilyn Monroe, The Seven Year Itch. Photo by Associated Press. No known copyright restrictions.

Well, opinions could legitimately differ. One person could say yes, and another viewer might say no. Without elaboration, the debate could very quickly devolve down to a "Yes, it is," "No, it is not" squabble.

The advantage of science is that the researchers have to spell out their reasoning. Let us take some of the reasons why the picture above (see Figure 1.6) could be sexist. One, the woman (Marilyn Monroe) is presented as a highly sexualized being. Two, the picture is voyeuristic. The skirt is lifted to reveal the hidden (forbidden) in her attire. Three, the bystanders are male—looking at Marilyn with a male gaze, including—if you look closely in the background—a male photographer. If the researcher developed a coding scheme that said, "A picture will be classified as sexist if it has four of the following characteristics (see

below)," then that coding scheme would provide the readers with a detailed view of what "sexist" meant to the researcher.

Sample coding characteristics for determining sexism portrayals:

1. Is the woman in the picture presented primarily as a sexualized object?
2. Is the woman an object for males to gaze at?
3. Is the woman nude or partially clothed?
4. Is her facial expression sexual (a pout, a sultry gaze, fully lips, tongue extended)?

In looking at the picture, the coder could reasonably say, "Monroe is a sexualized object" (criteria 1). She is an object for males to gaze at (criteria 2). She is shown fully clothed, and she is in control of her skirt, pushing the skirt down with her hand and elbow, but she is—due to the wind—partially clothed (yes on criteria 3). She has no sexualized facial expressions, no exaggerated sexual pout, no sultry gaze, and no flushing (no on criteria 4.).

Reasonable (that is, sane) people would most likely agree that Monroe is partially clothed and she does not have a sexualized facial expression, but she is the subject of an observer's gaze and viewed in a voyeuristic way. Therefore, using the sample coding scheme, all people who see the picture would have to agree Monroe is *not* portrayed in a sexist way. However, reasonable people might still think that the image of Monroe is sexist *even though* they have coded the picture as not sexist using the researcher's code.

The people who think that the image of Monroe is in fact sexist *have* to (by the rules of science) agree that according to the methodology used the picture is not highly sexualized. However, they can argue that the method did not actually completely capture what can reasonably be defined as a sexist image. In other words, you turn a "Yes, it is," "No, it's not" moment to a deeper reflection on sexism; that is, "If I still think that this picture is sexist even though the method used classifies the image as not sexist, what criteria am I a*ctually* using to define sexist?" You still need to say that according to the criteria the scientist used, the picture is not sexualized, but you could say that the method is wrong, or incomplete, and according to the criteria that *you* develop, the picture *would be* sexualized. (For example, you might argue that pictures that have two of the four criteria are sexist, you might argue that the aspect of voyeurism alone is enough to classify the picture as sexist, or you might develop a whole new definition of sexism.)

You then have grounds to argue that the methodology of the study is incorrect, but you still have to accept that using the criteria that the researchers used the picture is sexist. Essentially, you have moved the argument to another "My method is right," and "Your method is wrong," which is simply another version of the "Yes it is," "No it's not" problem, *but* you have also clarified the meaning of what constitutes sexism to be used in another study.

Another study using the new methodology would have the same rules that accepting the methods means accepting the findings. If the new method suggests that the image is sexist, well, then, the two studies collectively mean that there is disagreement on what definition of sexism is acceptable. If the new method still does not classify the image as sexist, then it is even more likely that the image shouldn't be considered sexist, because two studies with different methodologies have come to the same conclusion. (Maybe what you consider sexist is wrong or maybe your definition still hasn't captured the specific things about the picture that indicate it is sexist to you.) In this way, what we mean when we use a concept increases in precision.

Science as a Way to Develop Policy: Science, Tobacco, and Climate Warming

Relying on science to understand the world around you (which is science's job, after all) is not the only way to develop an understanding of or form a plan to change the world. You can rely on luck, rely on what has worked in the past (tradition or custom), or apply a principle. However, relying on evidence gathering and data, which is the essence of the scientific approach, allows policy makers to understand and construct policies that cope rationally with the more subtle kinds of problems that spread throughout the nation (wide scope), penetrate into our cells (nonvisible effects), or take decades to develop (slow emergence). This approach intrinsically values facts and data over opinions (see Table 1.2), even when those facts are unpleasant.

Table 1.2. Differentiating Fact and Opinion

Differentiating Fact and Opinion		
	A FACT can be VERIFIED	**An OPINION is what one feels, thinks or believes**
Key Words	**Dates, historical events, numbers, science, data**	**Believe, feel, think, just know, am sure, some people say...**

Climate warming is already here—rising sea levels, wildfires, heat waves, coral reef die-offs, hurricanes, floods, virus/disease spreads, and mass species extinction. None of us can personally demonstrate global warming from experience. An older person can say that "there were more reefs when I was young," but to demonstrate that the death of reefs is related to cutting down

trees in the Amazon or cow gas (burps and farts)[9] is beyond what a single individual can reasonably determine from personal experience.

Instead, numerous people use agreed-upon procedures (that are open to inspection) to determine individual bits of information that make the connections between greenhouse gasses and methane, temperature increases, and coral reef death. Together, scientists agree that if they trust the methods that their fellow scientists are using, then they trust the findings. When study after study looks into different aspects of climate warming, and year after year the findings imply climate warming is/will be a problem, and further, when predictions scientists make on the impacts of climate warming repeatedly come true, then scientists develop a great deal of trust in the combined weight of these findings.

Our culture — and specifically our government — tends to put an enormous degree of trust in scientific findings. When scientific evidence becomes overwhelming, that evidence can be a counterweight to narrow interests profiting from a product or practice. The gradual development of a national consensus on the dangers of tobacco smoke is the classic case of the triumph of science and the public interest over the narrow interests of an industry.

At the end of the 19th century, both cigarette smoking and lung cancer were rare; cigarettes were considered "immoral" and "dangerous," and ready made cigarettes were a luxury good. The development of a cigarette rolling machine[10] and a "brilliant businessman who virtually invented mass marketing and mass distribution"[11] transformed cigarette smoking from an upper-class vice to a common habit.

In a fifty-year period between 1900 and 1950, cigarette smoking soared, and so did lung cancer. By the half mark of the century, the scientific evidence that smoking caused lung cancer was developing rapidly. The surgeon general's report in 1964 based on a review of over 7,000 published articles

9. Not a joke. Cow farts contribute to climate warming. Cows crop and chew grass and pass the plant down the esophagus to the first compartment of the cow's stomach, which can hold twenty-five or more gallons or more of material. The rumen acts as a fermentation vat that churns the food microbes within the rumen digest. The process produces thirty to fifty quarts of gas per hour, mostly carbon dioxide and methane, which the cow must release or die. James Linn et al., "The ruminant digestive system," University of Minnesota Extension, 2021, https://extension.umn.edu/dairy-nutrition/ruminant-digestive-system.
10. The development of the machine was a direct result of a $75,000 prize offered by a tobacco manufacturing company. Allan Brandt, *The Cigarette Century: The Rise, Fall, and Deadly Persistence of the Product that Defined America* (New York, NY: Basic Books, 2009).
11. During World War I, the military was giving cigarettes to the troops to "lighten the inevitable hardships of war." Brandt, *The Cigarette Century*, 104. In the decades just after the war, filmmakers glamorized smoking, while doctors and athletes were featured in ads praising the cigarette they smoked. Martha Gardner and Allan M. Brandt, "The Doctors' Choice is America's Choice: The Physician in US Cigarette Advertisements, 1930-1953," *American Journal of Public Health* 96, no. 2 (February 2006): 222-232, https://doi.org/10.2105/AJPH.2005.066654.

stated conclusively that smoking increased the risk of developing lung cancer. Neither government authorities nor university scientists have wavered from this conclusion.

The tobacco industry hired scientists, mostly scientists outside of the medical field, on their behalf. They also systematically funded and promoted studies that showed that secondhand smoke did not cause cardiovascular disease and that low-tar cigarettes are low-harm cigarettes (Tong and Glantz 2007);[12] however, the Big Tobacco playbook of challenging science has now lost the war on public trust,[13] and significant policy has been developed to reduce public smoking.[14] Science, credible and independently funded, had the power to stand up against extremely well-funded special interests, which makes science one of the few forces in society that can support the less powerful over corporate interests. To put the issue in the language of the populist movement of the 1900s, the "little guy" over "the man."

Disadvantages

Science is not the perfect weapon for social justice. Building scientific knowledge is slow and expensive, and inherently cautious. The benefits of science are most likely to go to the educated, who are more likely to find the information and to benefit from its content sooner. However, in those cases where science does demonstrate harm, science (combined with activists and the law) is a powerful weapon to support change.

Understanding the rules that science operates by.

The goal of this book is to review the basic rules of judging three of the most used methods in communication sciences—content analysis, survey research, and experimental methods. These are not the only social science methods, but they are widely used to study texts, people, and impacts.

First, all methods have rules that allow the informed reader to decide how much to trust the study's findings. These rules involve *appropriateness*, *bias*, and *verisimilitude*. *Appropriateness* translates into, "Is the method selected appropriate for the kinds of questions that the researcher

12. Elisa K. Tong and Stanton A. Glantz, "Tobacco Industry Efforts Undermining Evidence Linking Secondhand Smoke with Cardiovascular Disease," *Circulation* 116, no. 16 (October 16, 2007): 1845-1854, https://doi.org/10.1161/CIRCULATIONAHA.107.715888
13. As of July 2017, ninety-five percent of Americans felt that smoking was harmful. Only two percent felt that it was not. Eighty-nine percent felt that second-hand smoking was harmful.
14. Cigarette smoking is largely banned in enclosed workspaces (in twenty-five states), including bars and restaurants. Slightly over eighty percent of all Americans live with a ban on smoking at their workplaces, and/or restaurants and/or bars. American Nonsmokers' Rights Foundation.

is asking?" In the same way that you do not use a tape measure to check how hot the oven is, the kinds of questions that content analysis answers cannot be asked with survey or an experiment. *Bias* refers to the multiple ways that the researcher can construct a study that favors one answer over another. A content analysis study that coded all pictures of adult humans with beards as male and all others as females would be biased towards[15] women because all adult males who shaved would be coded as female. *Verisimilitude* (what scientists call validity) is the degree to which the specific group of things that are studied (the sample) are similar to the actual thing. All three are important to establish when the findings from a single study can be extended to a different population.

Second, all studies have limits. The findings are good for the study, but extending beyond the study is problematic. For example, a beautifully constructed sample of South Carolina residents will reflect the views of the population in South Carolina, but is not necessarily trustworthy for Illinois. If the topic of the study is, "What is proper outdoor clothing for winter?", then the two populations are likely to give very different answers. A careful reader should say that they trust the answers for South Carolinians, but not for Northerners. That does not mean that the study method was wrong; it means that a careful reader would recognize that the people studied (South Carolinians) are systematically different in important ways from the people for which the policy was developed (Illinois residents).[16]

Third, one study is one study—no more and no less. Because individual studies have limits, policymakers need to pay attention to how robust the scientific knowledge is. This is one of the major areas in which social scientists need to be more careful than physical scientists. In physical science, researchers (and research readers) generally assume that if someone drops a ball from a specific height and measures how many seconds pass before that ball hits the earth, then anyone can drop the ball from that distance, any place on Earth, any day of the year, and get the same results. Social scientists cannot assume that showing the same television show to any group of people will produce the same results, because the audience might view the show through different cultural lenses. In fact, scientists cannot assume that showing the same show to the same people on the same day will produce the same results (because the second showing loses the element of surprise). However, if several studies among different populations at different times have the same findings, we say that the findings are robust. Policy makers tend to trust robust findings.

The following chapters will cover rules for three major methods—survey, content analysis, and experiment. The final chapter looks at how to weave the findings (from sound studies) together to

15. "Biased toward" means that the researcher will overestimate in one direction. In this case, the coder will count more people as female than are female (females counted = females + beardless males).
16. For those who are not familiar with either South Carolina or Illinois weather, South Carolina is warm in January, and Illinois can get quite cold.

answer researchers' questions, whether their question is "What is the effect of Twitter on rational discourse?" or "Are Instagram fitness pictures self-sexualizing?" or some other question entirely.

PART II

SURVEY ANALYSIS

"Aw, people can come up with statistics to prove anything, Kent. Forty percent of all people know that."

—*Homer Simpson quote from The Simpsons episode "Homer the Vigilante"*

"I like to do my principal research in bars, where people are more likely to tell the truth or, at least, lie less convincingly than they do in briefings and books."

—*P.J. O'Rourke*

Survey Police[1]

1. Survey Police, "Famous People Talk about Market Research," Survey Police Blog, last updated November 3, 2022, https://www.surveypolice.com/blog/famous-people-talk-about-market-research/.

2. Judgment Rule 1 for Surveys

Judgment Rule: Check whether a survey method will adequately provide an answer to the research question.

Key Takeaways

Judgment rule answers the question: *Is survey research the appropriate method for answering the research question?*

Survey research is one of the most common methods of social science research, primarily because the method is a relatively simple and direct way to find out about people—you ask them.

Survey research is used to gather information when the research subject (the respondent) (1) can and (2) will answer a researcher's questions. When both of these basic conditions are met, the researcher and the reader can get a pretty good idea about what a person thinks, feels, or does by asking. If the researcher asks a number of people in a group—and those people adequately represent who is in the group—then the researcher can reliably describe what the group as a whole thinks, feels, or does.

To put more formally, social science survey research is a systematic way to ask a selected group of people "a question or a series of questions in order to gather information about what most people do or think about something", and to do so in a way that "examines and delineates the form" of the group.[1]

1. *Merriam-Webster*, s.v. "survey (*v.*)," accessed July 8, 2014, http://www.merriam-webster.com/dictionary/survey.

Start with the Research Question

Any research question that can be answered by asking people questions can appropriately be researched with a survey.

Surveys are good for asking questions about:

- **Values or beliefs:** Do you believe that college athletes should get a share of college television revenue? Do you believe that the college athletes should get a salary? Do you believe that nudity should be allowed on television? Do you believe that privacy is a right?
- **Feelings:** What feature do you like most about your iPhone? What is the primary satisfaction you get from using Twitter? Have you ever been scared watching a movie?
- **Knowledge:** Who do you think is the president of the United States? Who do you think is the richest person in the United States? What movie studios can you name? How many symptoms of breast cancer do women know about?
- **Actions or behaviors:** Do you use Twitter? Do you have a Facebook account? Have you seen a movie in a movie theater in the last week? Have you ever screamed in a movie theater (while watching a movie)? Have you ever screamed during a movie that you were watching at home?
- **Classification or demographics:** What party did you vote for in the last political election? What is your occupation? What is your income? What is your gender?

To determine whether a research question can be answered with survey data, the reader must first ask what the research question is, and then whether respondents can accurately answer questions about that thing that the researcher is studying. In most cases in social science research, the researcher's fundamental question is a simple reworking of the article title into a question. From the research question, the reader should be able to get a good sense of who the researcher should be talking to and whether the specific survey questions asked will allow researchers to answer their research question.

Definition Box 2.1: Definition of Research Question

The research question is the overall question the research project is trying to answer.

To illustrate: For the research article "Psychological well-being and demographic factors can mediate soundscape pleasantness and eventfulness,"[2] the research question is, "Does psychological well-being and demographic factors affect (mediate) the subject's perception of soundscape pleasantness and eventfulness?"[3] The article "Cameras of Merit or Engines of Inequality? College Ranking Systems and the Enrollment of Disadvantaged Students" is answering the question, "Do college ranking systems promote inequality between disadvantaged and advantaged students, or do these ranking systems help the truly meritorious (regardless of the student's background)?" Only in rare cases in peer reviewed journal articles, and more commonly in books,[4] will the title not communicate the research question.

The research question also will tell the reader who the researcher should have talked to (the research population) and will establish the basic criteria that the reader needs to know in order to judge whether a specific question will help provide an answer to the more general research question.

The research question, "What is Latino/Latinas' level of knowledge about diabetes prevention and treatment," for example, establishes that the researcher must talk to Latino and Latina participants and must ask questions that determine the survey group's level of knowledge about diabetes. In interviews, the researcher should ask specific questions about diabetes: "Does diabetes run in families?" (It does), or, "Are shortness of breath and chest pains symptoms of diabetes?" (No, those are the symptoms of a heart attack). Based on how people answer these questions, we can find out how much they know about the disease. We can also ask people where they got their information, which would be useful in answering the question "Where do Latinos get information about diabetes?" The interview questions could be fairly straightforward ("Where do you get this information?"), or the researcher could give people a range of choices ("Where did you learn that diabetes causes shortness of breath and chest pains: Family? Friends? Your doctor? The web? The library?")

The exact interview question, in a sense, does not matter (but see the section on judgment rule three). There are a lot of potential questions that could be asked that are part of the general universe of what is known about diabetes and a lot of ways to phrase a question about how the respondents got the information they have (or think they have). What is important is that the

2. Mercede Erfanian et al., "Psychological Well-Being and Demographic Factors can Mediate Soundscape Pleasantness and Eventfulness: A Large Sample Study," *Journal of Environmental Psychology* 77 (October 2021): 1-8, https://doi.org/10.1016/j.jenvp.2021.101660.
3. James Chu, "Cameras of Merit or Engines of Inequality? College Ranking Systems and the Enrollment of Disadvantaged Students," *American Journal of Sociology* 126, no. 6. (May 2021): 1307-1346, https://doi.org/10.1086/714916.
4. For example: The central thesis of *The Innovation Complex: Cities, Tech and the New Economy* (which tracks the complex relationship between a tech meritocracy, investors, elected officials, real estate developers and universities) is that these sectors of society are melded together in an alliance of self-interest. Sharon Zukin, *The Innovation Complex: Cities, Tech and the New Economy* (New York: Oxford University Press, 2020).

researcher could answer his/her fundamental research question by asking people questions, and therefore, survey is an appropriate method to use. (For more examples, see below.)

Definition Box 2.2: Definition of Interview Question

One of a series of questions asked to a survey respondent. The collected set of interview questions should answer the overall research question.

Additional Examples of Research Questions

- What bothers children online? This one is fairly simple.
  - Who to talk to: children
  - Sample question: "What have you seen on the internet last week that made you sad or unhappy?"
- Are depressed women who report that their health providers listen to them more likely to get help than depressed women who say their health providers do not listen to them?
  - Who to talk to: depressed women
  - Sample questions: Do you feel that your health provider listens to you? What help did you receive for your depression from your health care provider?

The answers to these two interview questions should allow the researcher to answer the research question. Survey is therefore an appropriate method.

3. Judgment Rule 2 for Surveys

Judgment Rule: Determine how the sample is different from the population.

Key Takeaways

Judgment rule answers the questions: *What population is the researcher studying? What group did the researcher sample? Did the sampling method introduce any bias?*

The second judgment rule is all about the difference between the population that the research is trying to find out about and the (usually) much smaller sampling of people the survey researcher has interviewed. When you are at a high-end ice cream shop and you get a sample of a particular flavor, in general, you assume that the small taste of blueberry walnut ice cream is going to be very much like (that is, the sample will represent) the flavor you would get from anywhere in the bucket. Your assumption is probably correct, but if the ice cream hasn't been mixed thoroughly, your sample might have a clump of nuts or no nuts at all, leading to an impression that isn't like the full blueberry-walnut ice cream experience. Unfortunately, in sampling a population, the nuts are often incompletely mixed, and so readers have to make special efforts to understand just how much trust they can have that the sample is just like the population.

Analyzing the research population: The reader's first task is to identify what population the researcher is interested in studying.

If the researcher is interested in the question, "What percentage of gamers are geeks?" then the population is gamers. If the question is, "What percentage of adults use Twitter?" then the population is adults who may or may not use Twitter (in other words, adults). If the question is, "What percentage of U.S. workers use computers as a part of their daily jobs?" then U.S. workers would be the population.

Of course, defining the population means providing clear definitions of who is a member of the population and who isn't. For example, does "workers" mean people who labor, or just the people who get paid for their labor? If the researchers are only looking at paid labor, they are leaving out

all people who stay at home and do housework/childcare/eldercare and are not compensated for their jobs. That is fine if what the researchers are interested is paid labor, but not if they are interested in all labor. Similarly, is the population that the researchers are actually interested in people who labor within the geographical United States (which would include undocumented immigrants), but not American citizens who are working abroad (which would include a decent chunk of the military)? In other words, who exactly are "U.S. workers"? Essentially, when researchers decide who legitimately is a part of the population, they are also deciding who is important and who is not. By revealed preference, those groups that are important will be included; those groups that are not important should be left out *deliberately*, not because the researcher is too lazy to be clear about the definition of labor they are using.

The issue of generalization: Why sample a few individuals rather than interviewing everyone?

Definition Box 3.1: Definition of Sampling

Sampling is the process of selecting a few to represent the many.

Usually, different populations are large enough that interviewing every person in a group is too expensive and too time-consuming to do, particularly when there is a cheaper and more effective alternative available—selecting a few people to speak for the whole. If done correctly, the sample will accurately reflect the range and types of responses in the larger population. On the other hand, sampling also is one of two major ways errors can creep into surveys.

To illustrate, let's go through an example where we know the actual characteristics of the population. The research question is: What uses and types of gratification do Twitter users get from using Twitter? In Table 3.1 (below), we have the results for the full population.

From Table 3.1, we can see that Twitter users tweet to communicate with friends, to get political news, to follow trends for work, and to listen to celebrities. But if you look closely, older users are distinctly different from younger users, and male users are distinctly different from females. Older people are far more likely to get political news or follow trends for work, while younger people are more likely to use Twitter to talk to friends and follow celebrities. Males are more likely to get political news and far less likely to follow celebrity tweets than females.

Table 3.1. Uses and Gratifications from Twitter

Uses of Twitter	**Under 30**		**Over 30**	
	Males	Females	Males	Females
To talk to friends	50%	90%	25%	25%
Get political news	40%	20%	90%	40%
Follow trends for work	0%	0%	50%	40%
Follow celebrity tweets	20%	90%	0%	70%

If the researcher has a representative sample, then (most likely) the sample will show the same, or roughly similar, percentages as shown in Table 3.1 above for the full population. (See Definition Box 3.2 for definition of representative sample.)

Definition Box 3.2: Definition of Representative Sample

A representative sample where each and every person in the population has an equal chance of being selected.

On the other hand, a sampling method that doesn't allow each and every person to have an equal chance of being selected is likely to produce a sample that does *not* reflect the entire population. Let say that a researcher—a female researcher—used the following method to get her sample. She tweeted to her friends (mostly female and under the age of 30); she put an announcement on her Facebook page (viewed mostly by females under the age of 30); she then asked these people to retweet and repost. By and large, her friends will also be communicating with females who are predominantly under the age of 30. According to Table 3.1, females under age 30 are much more likely to use Twitter to talk to each other and to follow celebrities, so a sample that is accurate for the female-under-30 group is likely to widely overemphasize how much the population as a whole uses Twitter to talk to friends and to follow celebrities, simply because the sample is much more likely to sample young females than males under age 30, and males and females over age 30.

Box 3.1

Look for whether the sample was chosen randomly!

The best way to get a sample that reflects the population is to allow each member of the population to have an equal chance of being selected.

Doing sampling. It is relatively easy to *name* a research population (geeks, Twitterers, workers), but *finding* the population can be more difficult. *Naming* the research population simply means identifying the group of people that the researcher wants to study – whether those people are practicing doctors, farmers, Shetland Sheepdog (Sheltie) owners, breast cancer survivors, or people who have been exposed to mercury.

Finding the survey population depends on being able to locate and identify each person who is in the actual study population and to keep out each person who isn't. Depending on how you select the sample, you could either get a sample that looks a lot like the population, or a flawed sample that is systematically different in important ways. In terms of the original description of a survey as "delineating the form and the outline" of the population, the flawed survey would miss some of the population's distinctive features. It would be the same as having a topological map that did not include some of the hills or the valleys in the real region.

For the respondents to reflect the population, the sample respondents need to resemble the range of groups in the population. The best way to get a representative sample is to allow each member of the population to have an equal chance of being selected. The actual mathematics of why random sampling works are fairly complex, but fortunately selecting a random sample is relatively easy. Since humans tend to have built-in biases or systematic ways of simplifying selections, the best ways to select an unbiased sample are (a) to have a computer generate a list of random numbers and use those numbers to select the sample from an already compiled list of the population, or (b) to develop a sampling method that gives everyone in the population an equal chance of being selected.

The sampling frame method: The sampling frame is a list of all people who are known to be in the population. If a researcher was interested in doing a survey of practicing doctors, they could buy a list of all physicians who are licensed to practice. (In fact, there are businesses whose job it is to constantly update and monitor this list.[1]) Finding Sheltie owners is more problematic. The Ameri-

1. Providing an accurate and up-to-date list of the population is not an easy or inexpensive task. Look at the effort made to develop and maintain a list of practicing physicians (as described by a company brochure advertising the complete-

can Kennel Association keeps a list of all purebred Shetland Sheepdogs registered with the organization, but this list wouldn't include either mixed breed dog owners or purebred Sheltie owners who haven't bothered to register. In addition, it is somewhat unlikely that the American Kennel Association constantly updates their list, so the lists they have are likely to include an unknown number of animal owners who have lost or given up their pets.

In the case of the population of "people who have mercury in their blood," not only is there no available list, but the people in the population might not even know that they are in the population. (That is, they might not know that they have detectable levels of mercury in their bloodstream.)

Table 3.2. Comparison of Various Populations with Sampling Frames Available to Reach That Population

Population	**Sampling frame**
Practicing doctors	AMA membership list
Farmers	Farm Insurance lists Tax rolls for farms
High school students in Deadwood, Wyoming	High school records of Deadwood
Sheltie owners	No known sample frame
Breast cancer survivors	No known sample frame
People exposed to mercury	No known sample frame

Generating randomness with screening. When there is no list, researchers cannot use a sampling frame to generate a sample. Instead, researchers need to develop another way to determine the population. Often, the method they use is to generate a large list that is random, but will have both

ness of their list of physicians): **A physician mailing list that paints a full and up-to-date profile of each practicing physician** SK&A A Cegedim Company advertising brochure: "Research Center verifies every field of every record in its database every six months to maintain accurate physician mailing lists. Additionally, every updated profile is reflected in the end user database the very next day. This commitment to quality means your marketing success. "Each business day, SK&A's Research Associates call more than 10,000 medical offices, hospitals, pharmacies and other healthcare sites to update our databases, fully verifying an average of 2,750 sites per day. This continuous telephone-verification process means you get the highest quality, most responsive physician mailing list available. "Reach practicing physicians only—not researchers, professors or other non-prescribers—where they make business decisions [...] at their offices. Market to other medical office decision makers, such as nurse practitioners, physician assistants, office managers or medical directors." "Physician List," SK&A, archived January 4, 2016, https://web.archive.org/web/20160104232556/http://www.skainfo.com/physician-mailing-lists.php

people who are in the survey population and people who are not in the survey population. The researchers will talk to all people on the generated list and remove the people who aren't actually a part of the research population, commonly by starting the survey with a screening question that asks if the person answering the question is actually a member of the population. The paragraphs below give an example of how survey method and a screening question work.

> In a study of cellphone and landline users, PEWS researchers used "a combination of landline and cellular random digit dial (RDD) samples [...] to represent all adults in the United States who have access to either a landline or cellular telephone. Both samples were provided by Survey Sampling International. [...] Numbers for the landline sample were drawn with equal probabilities from active blocks (area code + exchange + two-digit block number) that contained three or more residential directory listings. The cellular sample was not list-assisted, but was drawn through a systematic sampling from dedicated wireless 100-blocks and shared service 100-blocks with no directory-listed landline numbers."
>
> Aaron Smith, Pew Research Center[2]

In other words, landline phone numbers have a pattern—some blocks of numbers are all businesses, some are residential and businesses combined, and some are all residential. Since businesses are *not* part of the desired population, then the blocks of numbers that were all businesses were deleted. The researchers then generated a random sample from all numbers assigned to the mixed business-residential block and then the residential blocks. Obviously, some of the numbers that the researchers generated from the mixed business-residential block would be businesses, and so the researchers would need to drop these phone numbers from the sample. Since the researchers don't actually know which numbers are for businesses and which are for homes, they will need to call everyone in the sample and ask if the number that they called is for a business or a home. Only those people who said the phones were their personal phone numbers continued on to answer the rest of the study. The people who said the number was for a business would be thanked (politeness counts), but not asked any survey questions. This process of making sure that the person contacted is truly a member of the survey population is called screening.

Checking randomness: Generally, opinions are linked to life experiences. In this society, as in many societies, our experiences systematically differ by major demographic variables—age, gender, sexual preference, occupation, race, family status (whether we have kids or not). In addition, many of our basic and derived values and attitudes differ by political orientation: Republicans and Democrats differ on lifestyle, energy consumption, and a host of issues outside of pure electoral politics.

2. Aaron Smith, "Americans and Text Messaging," Pew Research Center, September 19, 2011, https://www.pewresearch.org/internet/2011/09/19/americans-and-text-messaging/.

Since we know that demographics systematically affect attitudes, one easy method to check the randomness of the sample is to compare the sample with known characteristics of the population the sample was drawn from.

Race, gender, and age are fairly common measures to use because the population distribution of each of these is often well known. In Example 3.1, for example, the researchers were studying adults (above 18 years old) in the United States.[3] We know, from both the U.S. census and other studies, the ranges in age, gender, race, and income of the U.S. population, and so we can compare known population characteristics with the sample. In this case, the sample clearly oversampled men and under-sampled females, African Americans and people in the highest income bracket.[4] Comparing what you could expect from the average population figures and what the researchers reported, the researchers did not capture the demographics of the actual population, a point the researchers implicitly acknowledged by characterizing the sample as a convenience sample.

Example 3.1

Example of including demographics

The sample represented diverse demographic backgrounds. Among the respondents (n = 699), 55.1% were males and 43.6% were females. The respondents ranged in age from 18 to 84 years). Among them 61.5% were younger adults (18–44 years), 28.2% mature adults (45–64 years), and 8.3% older adults (65 years or older). The majority were whites (80.1%), followed by African Americans (7.2%) and Hispanics (5.0%). A total of 104 (14.8%) respondents reported they had been diagnosed with clinical depression. The majority (60.7%) of respondents had an annual household income of $25,000 to $99,999, with 26.3% reporting $25,000 or less and 10.3% reporting $100,000 or more.

Efforts were made to obtain a sample of Internet users representing a range of key demographic attributes of the U.S. population. However, due to unequal response rates from the demographic subsegments, the resulting sample of respondents diverged from the demographic characteristics of the U.S. population. In light of this issue and a low overall response rate of 1.72%, the sample should be characterized as a convenience sample.

3. Jin Seong Park, Ilwoo Ju, and Kenneth Unhan Kim, "Direct-to-Consumer Antidepressant Advertising and Consumers' Optimistic Bias about the Future Risk of Depression: The Moderating Role of Advertising Skepticism," *Health Communications.* 29, no. 6 (October 2014): 586-597, https://doi.org/10.1080/10410236.2013.785318.
4. The 2010 census report shows that roughly 49 percent of the population is male; the black or African American population is slightly over 13 percent of the population; and 80 percent of household incomes are over $101,582.

Authors of research studies are obligated by custom to include a description of how they sampled their population. While I'm sure that there is at least one case of a study slipping through the review process that doesn't include the sampling technique, in forty years of reading social science research, I have not run across an example yet – and I'm not expecting one, either. Commonly, however, researchers assume that readers know both the strengths and the limitations of different major types of sampling. The next section will review basic types of survey sampling.

Different Types of Survey Sampling

Simple random sampling: In this type of sampling, every person or unit in the sample has an equal chance of being selected, either by a list of random numbers or by electronically generated random numbers. For example, let's say that the University of Illinois housing office wants to find out what freshmen in the dorms think are the top three problems affecting dorm life.[5] The university housing office has a list of all people living in the dorms, so it would make sense for them to use the sampling frame method. In order to get a random sample, the researchers would give each person on that list a number (from 1 to the last person in the list). The researchers would then decide how many people they want to interview and generate a list of that many random numbers. The list would tell the researchers which people in the population list to interview. (That is, if the list of random numbers was 7, 23, 47, and 56, the researchers would skip the people in the population list who were assigned the numbers 1, 2, 3, 4, 5, and 6, interview the person in the population list who was assigned the number 7, skip the people who were assigned 8 through 22, interview the 23rd person, and so on.)

The quality of a simple random sampling is primarily determined by the quality of the sampling list and the sampling frame. In the example given above, the housing list is likely to be very accurate. Since the university housing office is getting a great deal of money from each student, they are likely to put a lot of effort into making sure that the list is both complete and current. Therefore, a random sample taken from the list is likely to be very good. One major flaw, however, is that the researchers might want to know more detail about some relatively small subpopulation, and a simple random sample might not capture enough people to draw accurate conclusions. In a population of 1001, with 800 reds, 200 yellow, and 1 blue, drawing a sample of 100 from that population would give some 80 reds (plus or minus a little random error), 20 yellows and 1 time out of 100 a blue. Nine-nine times out of 100, however, just by random chance, there would be no blue in the sample, and therefore all of the blue's opinions would be missed. This might not matter that much

5. Obviously, since freshmen are not the only class rank in a dorm, the survey would need to start with a screening question: "Are you a freshman?"

if what the housing authority was looking at was color preference for rooms, but if the population was 800 cis women, 200 cis men, and 1 trans woman, the housing authority needs to know the preferences of that one transgender student to make good (that is, inclusive) policy decisions.

Stratified sampling: Stratified sampling is a technique that researchers use when it is important to sample members of subgroups within a population. It's commonly used during elections when the pollsters want to make sure to include minority populations (such as racial or religious populations) in the same proportions as those groups occur in the actual population. Let's say that the pollsters wanted to check religious reaction to some proposed legislation. The US population breaks down into the following proportions of known religious preferences: Christian 70.6%, Jewish 1.9%, Buddhist .7%, Muslim .9%, Hindu .7%, Unitarians and liberal faiths 1%, New Age Spirituality 0.4%, and Native American 0.3%.[6] To take a stratified sample, the researchers would randomly sample the same proportions of respondents as in the general population, or—out of a thousand-person sample—706 Christians, 19 Jews, 7 Buddhists, 9 Muslims, 7 Hindus, 1 liberals/Unitarians, 4 New Age, and 3 Native Americans (religious, not necessarily ethnic, representatives), which in practice means that once they filled their quota for Christians, they would keep sampling for the other categories until they found the prescribed number of Muslims, Hindus, Unitarians, New Age and Native Americans. Again, this is an acceptable method of random sampling, although it means that the researchers will have to adjust statistically for this oversampling. (The statistical adjustments – while interesting — are outside the scope of this book.)

Cluster sampling: This type of sampling is useful when the research population has naturally occurring clusters, such as schools or families. The clusters should be relatively homogenous within each cluster (for example, all grade schools are divided up into grades or levels, families into adults and children, Dungeons and Dragons players into cleric, druid, fighter, and so on). In the cluster sampling method, the procedure is to randomly select among clusters and then exhaustively sample within the cluster. Examining research that uses this technique, you should first look at whether you have naturally occurring homogeneous categories. For example, you would have to determine if all the "groups" in a Dungeons and Dragons sampling were Dungeons and Dragons players, or whether some of the "groups" were actually—for example—teams of paintball players who were wandering through but not actually playing the D&D game.). Once the researcher selected a D&D group, however, he or she should interview all people in the group. The same goes for families; the researcher selects a random sample of families (this is the random selection among clusters), but within each family selected, the researcher should survey each member of that family. Again, this type of sampling is considered to produce representative random samples.

6. Pew Research Center, "Religious Landscape Survey," Pew Research Center, 2015, https://www.pewresearch.org/religion/religious-landscape-study/.

Convenience sampling: Convenience sampling literally means sampling whoever is convenient, including volunteer sampling. This kind of sampling is easy, but seriously biased. In some cases, when obtaining the population is extremely difficult—interviewing lesbian couples in a strongly anti-gay setting, for example—researchers will use convenience samples, but readers generally will have strong reservations about the generalizability of that survey. (The safest assumption to make is that the sample represents the opinions of some people in the population, but you cannot tell the range or distribution of the population.)

Snowball sampling: Snowball sampling is a technique used to reach difficult or very small subgroups of a sample. For example, snowball sampling is useful for reaching people who either:

1. have a rare characteristic (such as an infrequent disease or some other trait of interest), or
2. have reason to hide—either from social shame (such as pedophiles) or legal reasons (such as people who are fencing illegal materials).

Once the researchers gain the trust of one member of the research population, they ask that people to refer them to others in the group. Using the logic that "birds of a feather really do flock together," the researchers use their interviewers' contacts to reach other members of the population. By definition, then, a snowball sample is not a random sample. The most that one can take away from this sampling technique is that some people feel, act, and believe what the sample believes, but one cannot generalize the data to the general target population.

Table 3.3. Quick Reference Table for Representative Sampling by Type of Survey Sampling

Type of sampling	**Potential for developing a representative random sample**
Simple random sampling	Random
Stratified sampling	Random
Cluster sampling	Random
Convenience sampling	Not random
Snowball sampling	Not random

4. Judgment Rule 3A for Surveys

Judgment Rule: Determine if the survey questions' answers will collectively answer the research question.

Key Takeaways

Judgment rule answers the question: *Will the respondent's answers to individual questions, as a group, provide an answer to the research question?*

Before the survey is distributed to the sample members of the population, the researchers also need to construct a list of questions to ask the respondents.

These questions, when answered of course, collectively have to provide an answer to the researcher's initial question.

Definition Box 4.1: Definition of Survey Questionnaire

A survey questionnaire is the list of survey questions the respondents are asked.

If, for example, the researchers are interested in the question, "Does watching Marvel movies decrease concern for environmental sustainability?" there are a number of different ways that the researchers could ask the respondent, and all would be good as long as the questions asked satisfactorily answer the researcher's research question. The first option, of course, is simply asking directly.

For Questionnaire A, the answer to question one will tell what the respondents believe to be the effect of watching Marvel movies (see Example 4.1).[1] However, the question asked is a behavioral question: Does a specific type of movie impact behavior? The question assumes that a respondent can reliably tell that their behavior is/or is not affected, which they most likely would not be able to tell.

Example 4.1

Questionnaire A: Marvel and environmentalism (short version)

1. Do you believe that watching Marvel movies decreases your concern for the environment?
    a. Yes
    b. No

Questionnaire B takes a different approach (see Example 4.2 below). First, Questions 1 and 2 determine if the respondents have watched any of the more recent Marvel movies, or if they are more likely to watch serious art films (*The Dig* and *Nomadland*) or lighter fare (*The Boss Baby* and *Cruella*). As a result, the researcher can develop a more nuanced understanding of whether people watch clusters of film types, and they can identify and contrast people who watch, for example, art films, with people who watch Marvel. In terms of answering the first part of the research question, "identifying who watches and who does not watch Marvel films," the answers will tell whether or not the respondent has watched the more recent Marvel films, so the reader can say, "Yes, these questions will allow the research to at least partially answer the first part of the research question."

The second set of questions gives the researcher a measure of belief in the need for environmental change which is presumable related to environmental behaviors. (Those who believe that action can help and that there is a need for action are more likely to support environmental activity than those who either believe nothing can be done or that there is no need for environmental protection.)

1. A further problem is that respondents know that the answer they "should" give is that they do care about environmental sustainability, so the respondents are likely to slant their answers to the more socially acceptable answer, but even if they didn't, all you would know is whether the respondents felt that this particular genre of film did or did not decrease concern about environmental issues.

Example 4.2

Questionnaire B: Marvel and environmentalism

1. Have you watched any of the following movies? Please circle all that you have seen.

Black Panther	Yes	No
Spiderman: No Way Home	Yes	No
The Dig	Yes	No
Avengers: Endgame	Yes	No
Stillwater	Yes	No
Cruella	Yes	No
Nomadland	Yes	No
Eternals	Yes	No
Joe Bell	Yes	No
The Boss Baby	Yes	No

2. Please circle the number that most closely mirrors how strongly you agree or disagree with the following statements on environmental degradation, with 1 indicating Strongly Agree, 2 = Agree, 3 = Neither Agree nor Disagree, 4 = Disagree and 5 = Strongly Disagree.

Statement	SA (1)	A (2)	N (3)	D (4)	SD (5)
It poses a hazard to the whole world.	1	2	3	4	5
It is on the rise.	1	2	3	4	5
It is under control.	1	2	3	4	5
The U.S. has strong environmental pollution laws	1	2	3	4	5

It is already impacting life as we know it.	1	2	3	4	5
It will impact future generations on the planet.	1	2	3	4	5
It is causing extinction of flora and fauna.	1	2	3	4	5
There's not much that can be done to stop it.	1	2	3	4	5
It is the government's job to deal with it.	1	2	3	4	5
All of us have to contribute towards reducing it.	1	2	3	4	5

Please answer the following questions about yourself.

3. What is your preferred gender identification?

    a. Male
    b. Non-binary
    c. Female

4. What is your age?

    a. Under 18
    b. Between 18 and 22
    c. Over 22

5. Are you in college or of out of college?

    a. Graduated from or in college
    b. Not a graduate from or in college

6. In the next election, are you more likely to vote mostly for Republican or for Democratic candidates?

    a. Mostly Republican
    b. Mostly Democratic

The alternative, Questionnaire C lists all of Marvel movies produced between 2008 and 2021 (the first major arc of the Marvel Multiverse).

Interestingly enough, although Questionnaires B and C ask very different questions, the two aren't that different in terms of determining whether or not the respondents watch Marvel Movies. Questionnaire B will give more insight into whether the respondents watch other kinds of films, and Questionnaire C will tell the researcher just how much of a Marvel fanatic the survey respon-

dent is (see Example 4.3), but both can determine whether the respondents are Marvel movie watchers, which is the primary research question.

Example 4.3

Questionnaire C: Marvel and environmentalism

1. I do not watch Marvel Movies. (Please skip to question 3.)
2. Have you watched:

Black Widow	Yes	No
Captain America: The First Avenger	Yes	No
Captain Marvel	Yes	No
Iron Man	Yes	No
Iron Man 2	Yes	No
The Incredible Hulk	Yes	No
Thor	Yes	No
The Avengers	Yes	No
Thor: The Dark World	Yes	No
Iron Man 3	Yes	No
Guardians of the Galaxy	Yes	No
Guardians of the Galaxy 2	Yes	No
Avengers: Age of Ultron	Yes	No
Ant-Man	Yes	No
Captain America: Civil War	Yes	No
Spider-Man: Homecoming	Yes	No
Black Panther	Yes	No
Doctor Strange	Yes	No
Thor: Ragnarok	Yes	No
Ant-Man and the Wasp	Yes	No
Avengers: Infinity War	Yes	No
Avengers: Endgame	Yes	No
Spider-Man: Far from Home	Yes	No

3. We would like to ask you a few questions about you. Do you:

Recycle paper	Yes	No
Recycle batteries	Yes	No
Recycle smartphones	Yes	No
Recycle computer equipment	Yes	No
Drive a car to work	Yes	No
Shower instead of bathe to save water	Yes	No
Turn the tap water off while brushing teeth	Yes	No
Vote for candidates who support environmental regulation	Yes	No

Please answer the following questions about yourself.

1. What is your preferred gender identification?

    a. Male
    b. Non-binary
    c. Female

2. What is your age?

    a. Under 18
    b. Between 18 and 22
    c. Over 22

3. Are you in college or of out of college?

    a. Graduated from or in college
    b. Not a graduate from or in college

4. In the next election, are you more likely to vote mostly for Republican or for Democratic candidates?

    a. Mostly Republican
    b. Mostly Democratic

In terms of whether Questionnaire B or Questionnaire C's survey questions will elicit more complete or more accurate answers to the second half of the research question (Question 3 on both survey instruments), again, both questionnaires will work. Questionnaire B's questions are phrased in terms of concern about the environment, while Questionnaire C's questions will elicit what survey respondents do.

Both questionnaires, then, will answer the research question, because both will elicit some idea on Marvel watching compared to environmental attitudes. The survey will *not* be able to tell the researcher whether watching Marvel films will decrease concern for the environment—the researcher would need a different method (an experiment) to determine causality. What a survey *can* do, however, is show whether the survey respondents who watch a lot of Marvel films have less concern for the environment compared to those who primarily watch other kinds of films (Questionnaire B) or fewer Marvel films (Questionnaire C).

5. Judgment Rule 3B for Surveys

Judgment Rule: Throw out any survey questions that are biased, misleading, or too hard to answer.

Key Takeaways

Judgment rule answers the questions: *Are the questions clear, direct, and within the respondent's ability to answer? Are the questions biased or leading?*

The reader needs to judge more than whether the survey questions will answer the research question; specifically, the reader needs to examine the quality of each survey question. Each question should be one that the respondents ***can*** answer, that they would be ***willing*** to answer, and that does not ***bias*** or ***mislead*** *the respondent.*

Put more positively, the questions should be clear, simple to understand, and not too demanding.

What should the reader do when the questions are biased or misleading? The reader should discard any answers to unclear, biased, or misleading questions, but throwing out one question does not affect the survey's overall reliability. However, if there is a consistent pattern of poorly worded questions, then that pattern indicates that the researcher either lacks skill or is willing to mislead the respondent. The reader should use a pattern of biased or misleading questions to judge the entire survey. Questionnaires from special interest groups, for example, are particularly likely to be filled with questions where the "correct" answer is the answer that is closest to the positions that the special interest groups take. These surveys are literally worthless in terms of understanding the respondent's actual positions. (See the section on social desirability bias for more information.)

As a part of judging the research, the reader should carefully examine each of the sample questions that the researcher provides in the methods or results sections of their research report, in order to check for a pattern of hard-to-answer or biased questions. Example 5.1 provides a sample of common hard-to-answer questions.

Example 5.1

Categories of hard-to-answer questions

Too vague: Had any problems with your computer lately?

Not mutually exclusive response categories: What is your current age?

1. Below 18
2. Between 18—30
3. Between 30—50
4. Above 50

Unbalanced scales: How much did you like the movie *Fifty Shades of Grey*?

1. Extremely
2. Very much
3. Mostly enjoyed

Not covering all answer categories: What electronic media do you own?

1. Phone
2. Computer
3. Television

Using jargon and acronyms: Which processor are you more likely to purchase?

1. AMD A10-6700
2. AMD A8-7600
3. Intel Core i7-3960
4. Intel Core i7-4790K
5. Some other processor

Double-barreled question: Do you have a cell phone or a landline?

1. Yes
2. No

Too demanding on memory: What movies did you watch when you were six?

Hard-to-Answer Questions

Vague Questions

Questions should be clear, as short as possible, and easy to understand—both the question and the answer categories. The question "How good is your television?" has the advantages of being short, but is actually vague and therefore confusing. Is the survey asking the respondents to talk about the quality of their reception, or the quality of the shows they watch? Either understanding of the question is reasonable, which means that neither the researcher nor the reader knows which interpretation the respondent is making. Basically, then, the question is worthless.

Vague Response Categories

Response categories follow the same general rules that questions do. The response categories should be clear, direct, and not confusing or leading. In addition, all response categories should be mutually exclusive and balanced, and cover all the potential answer categories. For example, in Example 5.1, the question "What is your current age?" has two response categories a 30-year-old respondent could legitimately fill out (response 2: between 18 and 30, and response 3: between 30 and 50). It is not clear which category the respondent should choose. Given that some of the respondents will chose one and some will choose the other, the researcher will not be able to make any clear conclusions about whether 30-year-olds are more like the 18- to 29-year-old group or the 31- to 50-year-old group.

There is a problem of another sort with unbalanced scales where the scale is weighed in one direction (usually favorable). The list of answers to how much the viewer enjoyed the movie *Fifty Shades of Grey* at *worst* only allows the respondent to say that he or she "mostly enjoyed" the movie; there is no response category, for example, for "absolutely hated."[1]

1. Given that *Fifty Shades of Grey* had a 24% rating on the Rotten Tomatoes tomatometer, a *lot* of people absolutely hated the film.

Jargon and Acronyms

Both jargon (specialized words used by a particular profession or group) and acronyms (initial letters of organizations) are confusing and irritating to those who are not familiar with the terms and are likely to be answered with a guess. Of course, the effectiveness of jargon also depends on the specific population studied. If the research population is composed of computer hardware specialists who buy computer equipment for university students and professors, you might expect this audience to be up-to-date with the newest technology and terminology in the field, so referring to computer processers by number (e.g., Intel Core i7-4790K) should be understandable to that population. If the survey respondents, however, are average computer users, they probably would not know these terms.

In addition, questions need to be appropriate for the respondent's age, culture, and ability to read (literacy level).

Double Negatives

Avoid double negative questions, which are simply difficult for any respondent to understand. For example, instead of asking people to respond to the following double negative statement—"It is not true that video games are not violent"—the researcher should have asked the respondents their reaction to the direct statement "Video games are violent."

Double-Barreled Questions

Double-barreled questions are two questions in one: Do you like rap and rhythm-and-blues? The answer might be clear if the survey respondent either doesn't like both or does like both, but if the respondent likes one and doesn't like the other, answering the question is very difficult. For example, the double-barreled question in Example 5.1—"Do you have a cell phone or a landline?"—assumes that people have either a cell phone or a landline. How should the people who have both answer this question? If the survey respondent doesn't get angry and quit filling out the survey (which does happen), they will make an executive decision to privilege one answer (either what they have or what they don't have) and ignore the other. Since no one other than the respondent knows what decision-making rule the respondent has used, the answer is essentially useless to understand what is going on in the population.

Questions That Are Too Demanding on Memory

Again, the purpose of surveys is to find out what the population is thinking, feeling, or doing, which is only a good method when the respondent is capable of answering the question. Questions that are too demanding or require too much effort to read or respond to are likely to get a guess at best or a random response, neither of which will deliver an accurate picture.

In surveys, a respondent's willingness to please can produce a response that portrays them in a more positive or favorable manner than their actual response would. This phenomenon is called social desirability bias. Even in the fleeting, and often totally anonymous, social interaction of a survey, people want to please—and they can be very skilled at reading what researchers want from the way questions are phrased. Respondents tend to over-report pleasing behavior—such as over-estimating how much time they spend in community service or how much money they give to charity. On the other hand, they tend to underreport behaviors that they think others will disapprove of—such as how much pickup sex they have or how racist they are.

Questions to Trigger Social Desirability Bias

Leading Questions

Leading questions, whether asked by survey researchers or the police, are "questions" that tell the respondent what answer the researcher wants. It is very unlikely that a respondent wouldn't know that the preferred answer to the question, "Do you agree that the natural beauty of the Arctic National Wildlife Refuge should be preserved?" is "Yes, by all means let us save the National Wildlife Refuge." The people who say they don't want to preserve the Arctic National Wildlife are fighting against a natural desire to please; therefore, it is quite likely that they are reporting their true feelings. It is also quite likely that some of the respondents who agree with preserving the wilderness are just agreeing with the

researchers' obvious preference, and that actual support for preservation may be lower than the survey responses would indicate.

Halo Effect

Researchers can also inject bias into a survey by tying the questions to an obvious good or an obvious bad.

For example, let's examine the question:

"Do you agree with the heavy metal guitarist who bites the heads off of bats when he says, "I believe that playing first-person video shooting games decreases violence?"

The question ties together beliefs about the impacts of playing first-person videos to feelings about heavy metal music and feelings about people who bite the heads off of bats.

Respondents who hate heavy metal music or think that heavy metal itself promotes violence and mayhem would be more likely to disagree with the statement, "I believe that playing first-person video games decreases violence," than they would have if they had been asked with no reference to heavy metal (or bats). (Of course, the people who love heavy metal (or biting bat's heads off) would also be likely to be influenced in the opposite direction.)

The tendency to react to what is tied to a question, rather than the question itself, is called the halo effect, and as with leading questions, is likely to shift the overall direction of respondents' answers. Those who like what the question is tied to will be more likely to agree ("Do you agree with Jesus when he said…"), and those who dislike the tie ("Do you agree with Hitler…") will be more likely to disagree with the statement.

6. Judgment Rule 4 for Surveys

Judgment Rule: Hesitate to accept surveys with a response rate lower than 60.

Key Takeaways

Judgment rule answers the question: *Is the response rate adequate?*

Response rate. The response rate is the number of people who successfully fill out and complete a survey questionnaire out of all people invited to respond to the questionnaire. If you sent out a questionnaire to 1000 people and 100 people answered, you would have a response rate of 10 percent. (It does not matter if the population the sample was drawn from was a billion people or a thousand; the response rate would still be 10 percent.)

The response rate is an indicator that the researchers have gotten most of the variety of responses that are in the population. If the response rate is a large number (over 60 percent is the general rule of thumb), most likely the researcher has gotten a representative range of the potential answers. That is, the 40 percent of sample who did not respond are presumably a lot like those who did respond. The lower the response rate, the less confident the reader can be in assuming the responses represent the population. With a lower response rate, there are a *lot* of people who are not answering, and many of those people could be quite different from those who do answer.

If you don't know how many people were asked to respond to the survey, you cannot know what percentage of the people who saw the questionnaire chose to not answer it. By definition, then, if you cannot calculate a response rate, all you can say is that some people feel, think, or do something, but you cannot generalize to the study population as a group.

Over the last several decades, the number of people who are willing to answer surveys has gone down, particularly with the use of online survey research (which commonly has a response rate between 10 to 30 percent). As a result, statisticians have put considerable effort into investigating

just what percentage of people need to respond to a survey to guarantee a representative sample. The answer is that there is no clear, absolute cutoff point in terms of the number of responses per 100 people contacted. If significant groups refuse to answer the questions, a response rate above 60 might still give a biased result. Any rule that is developed about response rates is a guide rather than an absolute indication of the external validity of a survey. In general, however, a response rate over 60 percent indicates a fairly good representation, while response rates of 20 to 30 percent are far more likely to misrepresent some groups. There is a caveat, however. Researchers and readers are far more likely to start trusting the results when multiple surveys on the same topic yield highly similar results. Twenty surveys with a 10 percent response rate that give very similar answers to the questions are collectively probably fairly accurate, while one survey with a 10 percent response rate could be accurate or could be wildly off—there just isn't enough information to tell.

7. Summary of Judgment Rules for Survey Methods

Decision Rule Set for Research Method

If a survey is a reasonable method for answering the research question, go on to the following decision rules. If NO, throw out the study discussion and conclusions. The answer to question 2 must be YES in order to have any trust in the study at all.

Decision Rule Set for Sampling

1. The survey can only be generalized back to the research population that was sampled.
    a. You need to check each time the researcher narrows the sample. How was the sample selected? How good was the sampling frame?
    b. Did the researcher check for external validity? (Did the researcher check his/her sample against the general population? That is, did the researcher check whether the demographics (age, race, gender) are like the general population?)
    c. Is the response rate respectable (see also section on response rate)?
2. If there are no major flaws with the selection of the sample, then you can trust that the survey results–for good questions–can be generalized back to the research population.
3. If the survey sample is NOT representative of the population, then you will have a range of answers.
    a. The survey results can be generalized back to the population, with the exception of specific groups that were not sampled. THE RESULTS PROBABLY REFLECT THE POPULATION EXCEPT FOR THE GROUPS THAT WERE NOT INCLUDED.
    b. The survey results show that some people have this opinion (feeling, belief, behavior), but the results shouldn't be extended to the population. THE FINDINGS APPLY TO THE GROUP SAMPLED AND JUST THAT GROUP.
    c. The sampling procedure was so flawed that you have to wonder whether the people who answered the survey might actually be a little weird (if only an unusual group of people might have answered the survey). DO NOT TRUST!

Decision Rule Set for Questions

1. Accept the findings when the questions are clear, easy to answer, and not leading or biased.
2. Findings (or conclusions) that are based on flawed questions should be thrown out.
3. You can also use the questions to judge a researcher's competence. If there is a consistent pattern of poorly worded or biased questions, you should assume that the researcher was equally sloppy (or biased) in other parts of the survey.

Decision Rule Set for Response Rate

1. Accept when the response rate is over 60 percent.
2. Accept with reservations when the response rate is between 10 and 59 percent. (That is, the findings apply to the group sampled and only that group.)
3. Under 10 percent, highly questionable. (The people who answered may be systematically unusual.) (Can accept if multiple surveys with a 10 percent response rate all give similar results using different sampling techniques.)

PART III

CONTENT ANALYSIS

"Research means that you don't know, but are willing to find out."

—Charles F. Kettering

"In fact, the world needs more nerds."

—Ben Bernanke

"As in the case of grammar, the 'rules' by which a social system works are not natural. They are sociological constructions. Although systems can be elaborate, in comparison with other kinds of inferences, extrapolations are relatively simple."
—Klaus H. Krippendorff, Content Analysis: An Introduction to Its Methodology

8. Content Analysis Introduction

Content analysis is the study of cultural artifacts—materials humans have produced that have meaning. Suppose you want to study how many movie characters smoked in 1930s movies and compare that to how many movie characters smoke today. Or any of the following questions:

- How do television newscasts portray other countries (England versus India, for example)?
- Do female characters in strategy video games have unrealistic bodies (i.e. overdeveloped breasts and overly narrow waists)?
- How do the lyrics in video rap music treat drinking alcohol and using drugs?

To answer each of the above questions, the researcher should look at the content of (respectively) television news coverage of England and India, female characters in strategy video games, and rap music lyrics—all texts that some person or group has produced.

While there are different definitions of content analysis (see Definition Box 8.1), they all share the basic focus on analyzing products that humans have produced, rather than asking people what they think or feel about human-produced material. This particular distinction is absolute. Content analysis is always looking at content, not perceptions of content. Content analysis cannot determine impact—how the text will affect you (or any other reader), *or* what you or any other reader will think or feel about the text.

Definition Box 8.1: Definition of Content Analysis

Berelson (1952, pg. 18): Content analysis is a research technique for the objective, systematic, and quantitative description of the manifest content of communication.[1]

Krippendorff (1980, p. 21): Content analysis is a research technique for making replicable and valid inferences from data to their context.[2]

Neuendorf (2002, p. 10): Content analysis is a summarizing, quantitative analysis of messages that relies on the scientific method (including attention to objectivity, a priori design, reliability, valid-

1. Bernard Berelson, *Content Analysis in Communication Research* (Glencoe,IL: Free Press, 1952).
2. Klaus Krippendorff, *Content Analysis: An Introduction to its Methodology* (Beverly Hills: Sage Publications, 1980).

ity, generalizability, replicability, and hypothesis testing) and is not limited as to the types of variables that may be measured or the context in which the messages are created or presented.[3]

Content analysis is further divided into qualitative and quantitative analysis. Qualitative is generally focused on developing theory (inductions) and not just testing (deduction); requires skilled social researchers who have developed their analysis through careful (also sometimes called "deep") reading of the text; and often is looking for the hidden meaning of the text, rather than the open or surface reading of the text. Commonly, the researcher makes assumptions about how the message was produced and how audiences read the content. In other words, the researcher is willing to assume that they can tell (from reading the text) what the producer intends to say and what the audience takes away from the message.

Quantitative context analysis, which is the primary focus of this chapter, concentrates on manifest content. The quantitative researcher is concerned with understanding what everyone sees—the surface meaning. It is extremely important in quantitative content analysis to count the same thing in the same way.

To roughly illustrate the difference between the two approaches, a qualitative researcher looking at the four pictures of wedding couples in Figures 8.1A-D could argue that modern western portrayal of a wedding couple emphasizes both heteronormativity and the "couple" as the most fundamentally important unit to the exclusion of the group or the community, a theoretical argument that uncovers the hidden assumptions about what is important in weddings—the couples. The emphasis on couples then becomes a fundamental building block for a heteronormative view of relationships that excludes extended family and community. Usually, the researcher will point out specific illustrative examples of "coupledom," but will seldom review or summarize their entire database. The reader then assumes, but does not know, that the illustrative examples are typical of the entire database.

A quantitative researcher, on the other hand, would first analyze the wedding picture through coding specifics of the picture, such as gender (number of males and number of females), setting (natural, built, combination), and focus of attention (whether the couples are more focused on the group or on each other). Quantitative researchers would count the number of females (three pictures have one female, and one picture has seven) and males (three pictures have one male, and one picture has seven). Two pictures have natural (outdoor) settings, and two have indoor settings, and the focus of the center is the male-female couple. From those findings, the researchers could

3. Kimberly A. Neuendorf, *The Content Analysis Guidebook* (Thousand Oaks, CA: Sage Publications, 2002).

develop a theory that the pictures emphasize heteronormativity and that the couple was the most fundamental unit for this type of picture, but the findings themselves would be the quantitative description derived from the coding scheme.

Figure 8.1A. Wedding couple on black sand seashore photo by Asdrubal Luna, Unsplash, Unsplash License.

Figure 8.1B. Wedding couple surrounded by crowd photo by Jonathan Borba, Unsplash, Unsplash License.

Figure 8.1C. Wedding couple in church photo by Karsten Winegeart, Unsplash, Unsplash License.

Figure 8.1D. Wedding couple standing on bare rocks photo by Allison Heine, Unsplash, Unsplash License.

That is, to do a quantitative content analysis, researchers select a sample of what they want to study (bits of TV shows, videos, games, film), develop a set of counting rules based on what they are going to look for (see Example 8.1), and develop a numerical description of the data.

Since content analysis involves meaning, part of the difficulty in doing content analysis is the degree to which audience members (those who receive the message) uniformly interpret the meaning of a text or picture. Quantitative researchers are extremely concerned that they minimize any ambiguity inherent in alternative interpretations of what the audience is seeing. That is, they want to make sure that they can precisely define what is counted.

Example 8.1

What can you study with content analysis?

Virtually any kind of cultural artifact, including: Maps, movies, videos, video games, dolls, toys, rap videos, magazine advertisements, YouTube videos, records, signs, television advertisements, maps, pictures, cartoons, romance novels, pop-up ads, reality shows, crime dramas, plays, mystery stories, works of art, postcards, country music lyrics, cereal package covers, pesticide labels, and many others.

"Counting rules" simply refers to how researchers will instruct coders to count. For example, let's say you were doing a study of how many statues in Norway were trolls (see Figure 8.2). Your first task would be to figure out what characteristics "counted" as "troll" versus something else—an elf, for example. You would develop a series of coding rules that distinguish "trolls" from "elves." Do trolls have big noses (yes), do elves (no). Do trolls have large feet (yes), do elves (no), wide mouths (trolls yes/elves no), potbellies (trolls yes, elves no), and tufted hair (trolls yes/elves no). Given these counting rules, then, would you code the statue (see Figure 8.2) as a troll or an elf?

Presumably, the vast majority would agree that (given the coding rules above) Figure 8.2 shows a troll, not an elf.

Coding rules for personal judgments are far harder to define. For example, it is much more difficult to say whether the troll is cute or fun than it is to code the troll as having a large or a small nose.

Secondly, the coding rules themselves may not adequately capture the essence of what needs to be coded. Do the coding rules for "troll" as given above really distinguish a troll versus a non-troll, or does the definition come too close to the edges of what can be considered "trolldom"?

Figure 8.2. Picture of being with large nose, pot-belly, and white tufted hair

Finally, what calling someone a troll means is deeply culturally embedded. A tourist from the United States in Norway, with a relatively limited knowledge of Norse folk mythology and Norwegian history, will probably view trolls differently than the more complex view of a Norwegian who understands the trolls' place in Norwegian history and the culturally sensitive way that Norwegians have been disrespected as "trolls."[4]

As a reader, what is important is to figure out how the researcher instructed coders to code and whether those coding instructions are sufficient for the purpose at hand. At one end, researchers can be very specific about their codebook, as in the following example on newspaper coverage of climate change (see Table 8.1).[5]

In this example, researchers have explicitly outlined the code for each climate change variable. Was climate change mentioned (climate)? Did a news article take a stand on whether climate change was a fact, or did the news article mention sources that claimed climate change was a hoax (climate perspectives)? Did the article talk about climate change as happening here (special proximity cues) and now (temporal proximity cues)? For each of the main variables, the researchers also included subcategories that show differences within each category. So, for the variable "climate perspectives," the researchers/coders would code the articles as either denying or confirming climate change. Codebooks also generally list cues that signal each subcategory for the coder; for instance, "claiming that climate change is a government conspiracy or that climate change does not exist" should be coded with the subcode "denial," in the category "climate perspectives."

Researchers do not always provide such a clear roadmap. However, commonly, research articles will include enough information in their results section to determine how the coders categorized their major variables (for example, see Table 8.2). In the study of my personal use of social media,

4. Norway, which was a colony outpost of other Scandinavians (including the Danes) through much of its history, has a complex relationship with trolls. Trolls in Scandinavian folklore are commonly seen as big, hairy, and slow to act, but fierce when roused. Norway's various colonizers commonly referred to Norwegians as "trolls" as a way to highlight stereotypes of Norwegians as "slow" but "potentially violent" (and therefore needing the "guidance" of other peoples.) On the other hand, trolls are a distinct part of Scandinavian folk tradition, of which Norway, as a culture, is extremely proud.
5. Roberta Weiner et al., "Climate Change Coverage in the United States Media during the 2017 Hurricane Season: Implications for Climate Change Communication," *Climatic Change* 164, no. 3-4 (February 2021), https://doi.org/10.1007/s10584-021-03032-0.

the variables are implied by the results table, specifically types of social media and purposes (productive purposes versus entertainment).

Table 8.1. Codebook Used for the Analysis of Sampled Newspaper Articles. Roberta Weiner et al., "Climate Change Coverage in the United States Media during the 2017 Hurricane Season: Implications for Climate Change Communication," Climatic Change164, no. 3-4 (February 2021), https://doi.org/10.1007/s10584-021-03032-0, reproduced with permission from SNCNC.

Codebook Used for the Analysis of Sampled Newspaper Articles		
Code	**Subcode**	**Description**
Climate	Explicit Reference	Must include one or more of the following phrases: climate change, global warming, global change, changing climate, or warming climate
	Implicit Reference	Does not include any of the explicit phrases listed above, but does discuss the changing frequency and/or intensity of hazardous weather, and/or changing temperatures
Climate Perspectives	Denial	Any reference to climate change denial perspectives, such as media figures claiming climate change is a government conspiracy, or that climate change does not exist
	Consensus	Explicit statements that climate change *is* happening, that climate change is a fact, or that a consensus of scientists agree that it is happening
Spatial Proximity Cues	Proximal	The effects of climate change are nearby to a reader in the U.S.; references to climate change impacts that occur in the continental U.S.
	Distal	The effects of climate change are far away from a reader in the U.S.; references to climate change impacts that occur outside the continental U.S.
Temporal Proximity Cues	Proximal	The effects of climate change are happening now; includes any present- and past-tense descriptions of climate change
	Distal	The effects of climate change will happen in the future

All research articles, however, should include enough information in either the methods section, an appendix, or the results section, that a reader should be able to clearly distinguish the variables and the subcodes of the variables well enough to be confident that they could code samples similarly to the author.

Table 8.2. Analysis of Productive Versus Nonproductive Time on Social Media

Implied Codebook in a Results Table		
	Productive use of social media (in minutes), presented in % of ***n***	Unproductive use of social media (in minutes), presented in % of ***n***
…twitter (before musk)	75	34
…twitter (after musk)	18	39
…snapchat	6	26
…facebook	—	1
Other	1	1
n	174	176

Content Analysis

Content analysis researchers (and readers) implicitly agree to the same general contract discussed earlier for social science—a study is only as good as the method, the method should be clearly described, and findings should be accepted only to the degree that the method warrants. Generally speaking, this means focusing your attention in three specific areas—the sampling method, the coding scheme, and the intercoder reliability. The remainder of this section goes through each in turn, starting with the most basic:

1. What is the research question?
2. Is content analysis appropriate to answer that question?

9. Judgment Rule 1 for Content Analysis

Judgment Rule: Content analysis is valid for studying text, not people.

Key Takeaways

Judgment rule answers the question: *Is content analysis appropriate for answering the research question?*

The first question that a reader needs to ask is, "What question is the researcher asking?" The researcher's question then becomes the standard for judging the rest of the methodological procedure. Fortunately, the question is usually a simple rewording of the article title into question form.

For example, the article "Examining Diversity: A Content Analysis of Cancer Depictions on Prime time Scripted Television" answers the question, "What is the racial and ethnic heritage of television characters with cancer on prime time scripted dramas?"

The reader then will judge whether content analysis is the correct method to use to determine the researcher's question. Since the researcher is interested in the content within television shows, content analysis is an appropriate method to use. (Reminder note: Content analysis is designed to study artifacts that humans produce, including texts, pictures, and videos.)

In Example Box 9.1, the research team is examining the visuals and lyrics of rap music videos to find out the themes of the rap music (conflict-oriented versus community-oriented) and the features of the performers (Eurocentric versus Afrocentric). Therefore, reasoning backwards, the research question is: "What is the content of popular rap music videos visually and lyrically?" Since the research question is about the content of something that humans have produced, rather than the humans themselves, content analysis is appropriate.

Example 9.1

Title: Controversial rap themes, gender portrayals and skin tone distortion: A content analysis of rap music videos

Abstract: A content analysis of rap music videos aired on BET, MTV, and VH1 examined the occurrence of controversial themes, gender differences, and skin tone distortion. The results of this study found that current rap music videos have placed an emphasis on themes of materialism and misogyny. Additionally, men and women in the videos differ in their portrayal of these themes. Specifically, female characters are significantly more likely to appear as objects of sexuality. Men and women also differ in their appearance, with more African American females appearing to have Eurocentric features. Implications and suggestions for future research are discussed.

Kate Conrad, Travis L. Dixon, and Yuanyuan Zhang, "Controversial Rap Themes, Gender Portrayals and Skin Tone Distortion: A Content Analysis of Rap Music Videos," *Journal of Broadcasting & Electronic Media*, 53, no. 1 (March 2009): 134-156, https://doi.org/10.1080/08838150802643795.

Usually, the researcher is extremely clear about what he or she is studying and how they are studying it. The tricky part is that sometimes the researcher makes a mistake, and they use the wrong method to answer their research question. If you accept the most fundamental argument about data-driven science, then you—as the reader—are bound to reject the study's conclusions if the researchers use the wrong method to answer their research question.

One of the most famous examples of a researcher making conclusions that go far beyond what his method can support is Lasswell's study of military propaganda, one of the classic studies of propaganda.[1] Lasswell, who studied the content of war material that the military developed for World War I (see Figure 9.1), found that military posters demonized the enemy (describing the enemy as "menaces" and "mad brutes"), appealed to fear ("loose lips sink ships"), and in other ways portrayed "our" side as good and noble and "their" side as animal and evil. Lasswell concluded from his study of the poster's content that propaganda built a base of support for the war, sapped the enemy's will to fight, is essential for modern warfare, and acted as a "flame to burn out the cancer of dissent …"

Since Lasswell studied posters, not people, he couldn't—scientifically—make any claims about the posters' impact on viewers. That is, he could demonstrate what kinds of claims the posters made, but not how the posters changed viewers.

1. Harold D. Lasswell, *Propaganda Technique in World War I* (Cambridge, MA: MIT Press, 1971), 268.

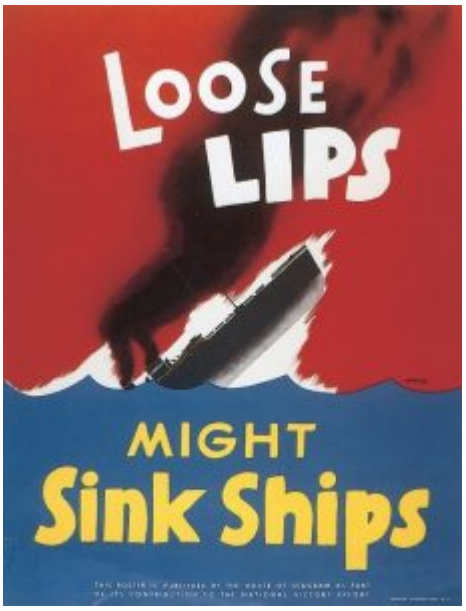


Figure 9.1. Poster of a liner sinking into the ocean with the caption "Loose Lips Might Sink Ships." World War I military poster by Seymour Goff, National Archives and Records Administration, public domain.

Figure 9.2. Poster of a bloody hand and forearm that is holding a blood-stained knife emerging from the sea. The poster's caption reads, "Help Crush the Menace of the Seas. Buy Liberty Bonds." J.L. Grosse, National Archives and Records Administration, public domain.

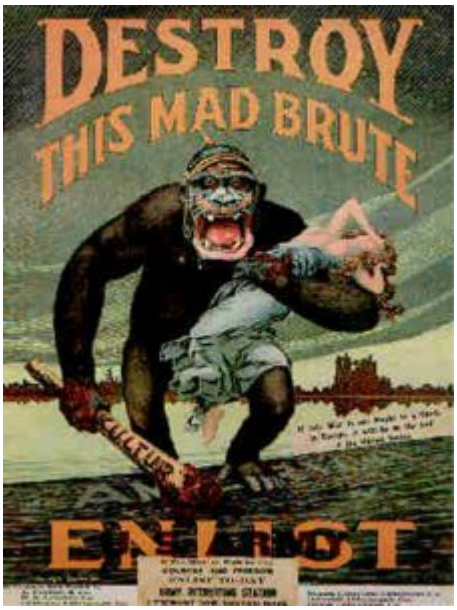


Figure 9.3. Poster with a gorilla-like beast with a German WWI military helmet abducting a obvious distressed maiden. The caption is "Destroy this mad brute. Enlist." Henry R. Hopps, Library of Congress, public domain.

When the researchers use the wrong method to answer their research questions, then their findings are useless for that research question; when the method they use is irretrievably flawed, then the findings are useless.

One additional note: The reader's "job" is to carefully distinguish between what researchers found (the methodologically sound findings) and what the researchers suggest the implications of those findings might be (in the conclusion or discussion section of the paper). Readers should accept methodologically sound findings as valid, but not automatically accept the researcher's discussion of the findings. In the discussion section of papers, researchers commonly suggest causes or implications of the findings. Both causes and implications are extensions beyond the actual findings and need to be studied further.

10. Judgment Rule 2 for Content Analysis

Judgment Rule: Determine how the sampling procedure changes the sample from the original population and how the changes from population to sample limit what conclusions can be drawn about the population.

Key Takeaways

Judgment rule answers the question: *What limits does the sampling method put on what can be said about the population?*

Subsections of Judgment Rule: *What is the population? What is the sampling frame (or sampling method)? How does the sampling frame differ from the population?*

Population

The population is the group that the researcher is interested in studying. If the research question is "How violent are video games?" then video games are the study population. If the research question is "Are toy dolls (action figures) for young boys hypersexualized?" then toy action figures for young boys are the population. As with survey methodology, the population of the study includes all those that are the objects of the study and only those.

Figuring out what is a member of the population and what is not a member may not be as easy to break down as it seems. For example, are Batman, Captain America, and the Teenage Mutant Ninja Turtles all examples of action figures for boys? What makes an action figure an action figure and not a doll? Is Queen Elsa from Frozen an "action figure" or a "doll"? Would Queen Elsa be tossed out of the action figure category because she is a girl-doll, and boys do not play with girl-dolls, or would she be tossed out because she is not active enough for "action" figure? If we take the definition of action figure to be "a doll representing a person or fictional character known for vigorous

action, such as a soldier or superhero," are Teenage Mutant Turtles, also known for their association with pizza, active enough to be considered action figures?

In describing a research study, the researcher should provide clear rules that distinguish what is included in the research population and what is not—both to increase clarity (as in the action figure example above) and to reduce an unmanageable question to a more easily researched focus. If the researcher was interested in knowing whether women are hypersexualized in media, then the population is all women, in all media, in all countries, in all time periods and all languages (an overwhelming task). In most cases, the researcher will narrow the study population to a much smaller group—for example, female characters in the top ten blockbuster movies in this century.

In each of the three examples (see Example 10.1), the researcher is narrowing the research population—Disney cartoons (Study 1), prime time dramas (Study 2), magazine ads (Study 3)—down from the theoretically interesting question "Are women hypersexualized in the media?" to a more easily researched question. Since the researcher has narrowed the research population, neither the researcher or the reader (you) can legitimately extend the findings beyond the research population studied. If the researcher did all three studies and all showed that women were hypersexualized, then you *still* could not answer the theoretically interesting question of whether all media hypersexualize women, but you could have a lot more confidence that large segments of the commercialized mass-produced media do.

Example 10.1

Narrowing the research question from "Are women hypersexualized?" to:

Study 1: Do Disney female cartoons hypersexualize female lead characters?

Study 2: Do television prime time dramas hypersexualize female characters?

Study 3: Do magazine ads aimed at early adult females hypersexualize women?

Let's go through one, more complex, example. In the study below, the researcher is interested in issues of race, specifically how many video games allow players to choose the race of their avatar. What the researcher is examining is whether video game players can construct an avatar with African features (head and hair). The researcher is also going to insert these findings into a complex argument based on past research. Research has shown that having a same-sex character in a movie or television show is important to viewers. Films generally have a range of characters, any one of which the viewer can identify with. Video games are different. Most research

suggests that game players exclusively identify with their game avatar. A female video game player, for example, will identify with her avatar, even if the avatar is male. If games systematically exclude some characteristic important for self-identification (such as race), then readers would conclusively know that video games were not, as a group, providing adequate positive self-identification for an important group. The overall argument is based on combining the findings from past studies with the specific findings from this study. However, you, the reader, are going to specifically be looking at whether the findings of *this particular study* are based on sound methods.

In terms of judging soundness, you need to ask specific questions about population. First, what is the population? (Answer: Avatars of role-playing video games.) Next, since the researcher is not going to look at all video games with avatars, how did the researcher select the specific games to look at? To answer these questions, you need to look at how the researcher described the method used to locate the games he studied (see Box 10.2).

Example 10.2

Avatars of Whiteness: Racial Expression in Video Game Characters

I set out to conduct a comprehensive survey of the character creation capabilities of both online and offline RPGs. I examined all RPGs released for the Windows computing platform in the United States from 2000 to early 2010. I chose this ten-year period due primarily to technical factors: identification of racial features on avatars requires a certain amount of graphical clarity which was generally not present prior to 2000, and avatar customization became more widespread with the advent of 3D graphics technology that allowed developer to more easily modify avatars than with traditional hand-drawn 2D graphics. Unfortunately, there is no authoritative listing of all RPGs produced in this time period. Consequently, I constructed a list from two online sources: IGN (pc.ign.com), a major gaming news Web site founded in 1996, providing hundreds of reviews of games of all platforms and genres and MobyGames (www.mobygames.com), which describes its mission as, "To meticulously catalog all relevant information about electronic games (computer, console, and arcade) on a game-by-game basis" using a Wiki-like model that allows users to contribute new information and correct the existing entries. From these lists, I searched for games within the RPG genre released for the Windows platform in the United States, which included "RPG" and "Action RPG" for the IGN database and "RPG" for the MobyGames database. The total number of RPGs from both sites was several hundred. However, MobyGames indexes not only original game releases, but also re-releases and expansions (add-ons to original games that are not capable of running alone). Thus, I eliminated all but the original releases of such games. Additionally, some games were eliminated in the course of testing due to technical issues where the games would not run on modern systems.

In addition to offline RPGs, I examined all MMORPGs operating and accepting U.S. players in the spring and summer of 2009. Again, there is no authoritative listing of operating MMOs. Consequently, I constructed a list from three online sources: Gamespot (www.gamespot.com), a major online gaming Web site that provides news and reviews of games; MMOGData (mmogdata.voig.com), a Web site dedicated to providing subscription data for MMOs; and OnRPG.com, a Web site that advertises itself as "the biggest and best Free MMORPG Directory on the net." Using information from all three sites, I compiled an initial list of 220 actively operating MMORPGs, including both games that have monthly fees and those that allow play for free. The total number was reduced in the course of downloading or attempting to purchase play time for these games, as a number were either no longer in operation, had Web sites or servers that were not operational, or had clients that would not run.

Additionally, to determine the capabilities of RPGs for creating non-white player characters, I eliminated those games that did not have a visible, human avatar (for example, games where you only controlled a vehicle or games without graphical interfaces). I also eliminated those games where the racial characteristics of the avatar could not be determined, for example, due to size/quality of the graphical presentation. Third, I focused only on those games that allowed character customization to see what constraints were placed on players who attempted to create a non-white virtual representation. After eliminating operating games that did not meet the above criteria, the sample totaled 20 offline RPGs and 65 MMORGs.

David R. Dietrich, "Avatars of Whiteness: Racial Expression in Video Game Characters," *Sociological Inquiry* 83, no. 1 (February 2013): 82-105, https://doi.org/10.1111/soin.12001.

Did the researcher use a reasonable method to a) develop the population and b) draw the sample from the original population of "role-playing avatar games?" I would say yes. He did a good job of finding role-playing video games for both online and offline games. There are some holes with the population (games that were no longer in operation, games that had poor graphics), so the research sample did not have a perfect population of all games released during his study period. However, we can be fairly confident that looking at five websites whose purpose is to gather and publish reviews of games will develop a list of most of the widely played and widely known role-playing games. All in all, the research did a great job of finding the population, with the caveat that he did introduce distortion by only using Windows-compatible games.[1] Since he did not give any information on whether Windows-compatible games are different from other computer games, this is potentially a serious distortion of the overall population of games and the reader needs to keep this distortion in mind when making any claims about games overall.

1. There were no errors introduced in sampling, because the researcher studied every game he identified. That is, he did a census (looking at the entire population), rather than a sampling (selecting a representative group from the population.)

Readers should always check whether the sample list that the researcher used actually reflects the population. Errors in this step are not uncommon. Sparkman's (1996) study of 163 U.S.-based advertising content analyses, for example, found that 97 percent of the articles claiming to represent national print ads were in fact samples that also included regional and local ads. Since researchers have also found that regional and local ads systematically differ from national ads on important characteristics of the ad (type of product, localization information, etc.), using generalizations from regional and local ads to make inferences about the character of national ads can be seriously misleading.[2]

Unfortunately, while researchers are expected to describe how they collected their samples, the conventions of research article writing do not force researchers to describe in detail how their sample doesn't resemble their research population, which forces the burden on the reader to think through how each step of collecting a sample introduces blind spots that ignore distinct subgroups in the overall population.

Example 10.3: Examples of Available Archives

CLIO Awards Archives

"CLIO Archives lists the cream of global advertising back to 1960." This site has a listing of winners from 1960 to the present, complete with info about each.

Lexis-Nexis

Lexis-Nexis is "the world's largest provider of credible, in-depth information," according to the company website. The Lexis database contains more than 22,000 sources, including full text archives of most popular newspapers and magazines.

Television News Archive—Vanderbilt University

The TV news archive at Vanderbilt "is the world's most extensive and complete collection of television news," according to the archive website. It has all network news broadcasts from 1968 to present.

Society to Preserve and Encourage Radio Drama, Variety and Comedy (SPERDVAC)

2. Richard Sparkman, "Regional Geography, the Overlooked Sampling Variable in Advertising Content Analysis," *Journal of Current Issues and Research in Advertising* 18, no. 2 (Fall 1996): 53-57, https://doi.org/10.1080/10641734.1996.10505051.

The SPERDVAC libraries contain over 2000 reels of old time radio, including *Hopalong Cassidy*, *Quiet Please*, *Mysterious Traveler*, *The Whistler*, *Dimension X*, *X Minus One*, *Tales of the Texas Rangers*, and *Inner Sanctum Mysteries*.

Steven Spielberg Jewish Film Archive

"The collection includes Holocaust films, films depicting Jewish life around the world, newsreels from Israel, and works from Jewish filmmakers."

Abstracted from Kimberly A. Neuendorf, *The Content Analysis Guidebook*, Sage Publications, Inc., 2017.

Sampling

The same types of sampling—census, random sampling, cluster sampling—used in survey analysis are also used for content analysis. The reader needs to check that the researchers gave each and every member of the population an equal chance to be selected, and, if not, consider how the systematic exclusion misses data that might be vital to fully understanding the research population. Each of the different types of sampling have their own limitations, discussed in the survey chapter, and the same limitations apply when used in selecting a sample for content analysis. However, content analysis has a few unique elements that are important.

Content analysis researchers are generally more likely to use archives, places where content has been stored (see Example 10.3 for a sample listing of available archives).[3] Kimberly A. Neuendorf, author of *The Content Analysis Guidebook*, includes the name, the website, and a short description of various online archives in the section on Message Units and Sampling: Archives. These archives are generally only as good as the procedures used to collect, index, and store the archived information. The first time you (and any other reader) see an archive mentioned, you should look up and understand how that archive was developed. The degree to which you can generalize the findings from using an archive is limited to how systematically that archive gathered its material. Vanderbilt television news archives and LEXUS-NEXUS, a news archive, are regularly used, and their limitations are well known. (When developing your basic research reading skills, you should read up on these archives the first time that you see them, but you will find many of these data sets, including LEXUS-NEXUS, are used time and time again.)

At the low end of generalizability, you will find personal collections such as the Gish Film Theater Collection at Bowling Green State University, which has donations of private papers from several

3.

Hollywood actors, including collections from two major silent film stars, sisters Dorothy and Lillian Gish. Private collections of papers are most likely to be collections of documents that the donors happened to keep (including correspondences). Embarrassing papers or the papers that the collectors considered "trivial" are likely to have been thrown out, which means that the archive likely presents a far more positive view of the object of the collection than would be presented using a different set of data.

Second, unlike surveys, many content analysis samples are selected from a population of the most popular examples of a media—the best-selling songs for various musical genres (often data gathered by Billboard), the most popular movies in terms of box office success (usually using the theater receipts from Box Office Mojo), the top prime time crime dramas on television (data usually derived from the Nielson ratings). In this case, the sampling is—obviously—representative of popular choices, but not representative of the total of what shows or songs could be watched or heard.

The reader, however, needs to keep in mind that audiences can self-select in very different patterns that could be critical for some research purposes. Would, for example, people with a low body image be likely to select a different media pattern than people with a normal or high body image? Are people who are troubled (in a variety of ways) systematically likely to view violent pornography or specific types of horror shows differently than people who are not troubled in the specific way under study? If so, then conclusions about the content of popular media selections would not necessarily apply to those specific groups of people, because these groups might systematically seek out a less popular set of viewing or reading material, or they could select just some specific types of content within the universe studied. Again, the same basic premise that was discussed for reading survey research also applies to reading content analysis research. The reader needs to carefully consider what exact population the sample is drawn from and not extend the findings beyond this group and the content studied, but also—if the reader is interested in applying the findings to the viewing or listening habits of a specific population of humans, they need to consider whether the population they are interested in has the same reading or viewing habits as those people who watch, listen to, or read the artifacts studied in the content analysis.

To illustrate, let's say that you were developing a set of recommendations for parents of high school children with attention deficit disorder on teen use of social media and smartphones. It turns out that teens with ADHD are not affected in the same way as non-ADHD or depressed teens, so in order to make recommendations about what kinds of use parents of children with these specific needs need to limit, the studies used to develop recommendations should be specifically studying smartphone use by teens with ADHD. The same reasoning holds for specific kinds of content. Horror content is, by definition, different from sitcom content. And using a more general sample such as "prime time shows'" depiction of on-screen violence might not give an accurate reflection of how much violent content would be seen by a group that is specifically looking for horror in the shows they watch. So, again, a careful reader looking for the answer to the ques-

tion "How much violence is my audience of interest (for example, depressed teens) exposed to?" would need to carefully delineate just how much they can abstract information from a content analysis of "prime time shows," given that the audience of depressed teens can systematically select for more horror (or less horror) in the shows they watch. That is, when looking through research papers to answer a specific question about a specific audience, the reader needs to keep in mind whether *their* population of interest is the same as the audience who is viewing the content studied in the content analysis study.

11. Judgment Rule 3 for Content Analysis

Judgment Rule: Coding categories should precisely measure the variable the researcher is studying.

Key Takeaways

Judgment rule answers the question: *Are the coding categories clear, easy to code, not leading or biased, and appropriate to answer the research question?*

Coding, in research terms, is where the rubber meets the road. If the theoretical framework is the map (the framework you use to tell you where you are going), and the road is the reality you are moving along, then coding rules are the wheels—the portion of the car that makes contact with what is "out there" and propels the research project (the car body) forward.

In all content analysis studies, researchers ask questions about the text. To answer these questions, the researchers need to describe precisely how they are going to measure each variable or characteristic of interest. To answer the question—*Do female characters in strategy video games have unrealistic bodies—overdeveloped breasts and overly narrow waists*?—researchers must determine precisely what is an "overdeveloped breast" and an "unrealistic narrow waist," and describe each in enough detail that trained coders "from varied backgrounds and orientations will generally agree in its application."[1] Coders follow researchers' instructions on how to code features of interest in the sample texts. (The set of instructions are called coding rules. Coding is the act of applying the rule to analyze a bit of text, video, or audio material.)

For example, how would a researcher develop coding rules to study Figures 11.1A-C (See Example 11.1B)? First, the researcher would need to have a research question such as, "Are women in magazine pictures portrayed as professionals less often than men?" Second, the researcher would need to develop a set of rules for all the variables needed to answer the research question. In the example below, the two major characteristics that the coder would need to record are "gender" and "

1. Kimberly A. Neuendorf, *The Content Analysis Guidebook* (Thousand Oaks, CA: Sage Publications, 2002), 9.

professional occupation." Coding rules that could capture these two characteristics are shown in Example 11.1A below.

Example 11.1A

The following is a sample of possible coding instructions for the figures in Example 11.1B, with the research question "Are women in magazine pictures portrayed as professionals less often than men?"

1. For the variable "Gender," code each person in the picture as either:
    - Female
    - Male
    - Unclassifiable
2. Occupation: Code each person in the picture as either:
    - Shown at work (as having an occupation)
    - Shown at leisure

The next step is finding material to code. An illustrative sample of pictures is shown in Example 11.1B.

Once the researcher has developed the coding instructions and found the sample material, coders enter their analysis of the pictures into a coding answer form. (See Example 11.1C for an example of a blank sheet and Example 11.1D for an example of a coded answer form using Figures 11.1A through 11.C from Example 11.B.)

For the pictures in Figures 11.1A-C, their coded data (see Example 11.1D) shows no men, and seven women. More women are shown at leisure (5 women out of 7 total), but more pictures show women working (2 pictures out of 3).

When all of the material is coded, then the researchers will tabulate the results, showing summaries that illustrate their findings (Table 11.3). Now the research can say with confidence that the women dominated the coverage (100%), although the majority of the characters were at play (71%) rather than at work (29%).

Example 11.1B

The following figures are sample material to be coded using the instructions in Example 11.1A.

Figure 11.1A. Picture of Sally Ride, America's first woman astronaut, from the U.S. Information Agency. National Archives and Records Administration, public domain.

Figure 11.1B. Picture of five Irish girls at the entrance to Tomorrowland, a dance festival in Boom, Belgium, by Eddy Van 3000, Flickr, licensed under CC BY-SA 2.0.

Figure 11.1C. Woman standing beside a Royal Typewriter, holding typewriter ribbon. Photo by Robert Yarnall Richie. Southern Methodist University, DeGolyer Library via Flickr. No known copyright restrictions.

Obviously, three pictures are too small a sample to draw any firm conclusions. If, however, after going over hundreds of pictures, the researchers found that 80 percent of pictures showed men at work and 90 percent showed women at work, then—clearly—women in magazines are shown as workers more often than men, albeit by a narrow margin.

Example 11.1C

Table 11.1 shows a sample blank coding sheet for coding the figures in Example 11.1B using the instructions in Example 11.1A.

Table 11.1. Sample Blank Coding Sheet

Sample ID	Gender	Occupation
1		
2		
3		

Coding for "gender" is fairly straightforward. Producers of mainstream media are generally simplistic in how they assign gender, and most audience members within the culture are fairly good at figuring out the codes that the producers use to signal gender.

Notice, however, that even with characteristics that are fairly easy to code, coders might have difficulty in some situations. The very old and the very young (Figure 11.2) are often not visibly gendered, unless the message producer uses conventional signals (dressing a baby in blue, tying a head scarf on a female) to signify gender. In addition, the producer could deliberately work to blur gender lines (see Figure 12.1 in the next chapter).

By definition, an adequate coding rule should be precise enough that two coders would code the same object the same way most of the time.

Example 11.1D

Table 11.2 is an example of a coding sheet filled out for coding the figures in Example 11.1B using the instructions in Example 11.1A.

Table 11.2. Coding Sheet Recorded Results for Figures 11.1A-C

Coding Sheet Recorded Results for Figures 11.1A-C				
Identification	Gender		Professional Portrayal	
	Female	Male	At Work	At Play
Photo 1: Person 1	1	0	1	0
Photo 2: Person 1	1	0	0	1
Photo 2: Person 2	1	0	0	1
Photo 2: Person 3	1	0	0	1
Photo 2: Person 4	1	0	0	1
Photo 2: Person 5	1	0	0	1
Photo 3: Person 1	1	0	1	0
TOTAL	7	0	2	5

Coding instructions range from unelaborated instructions, such as "Record hair color," to highly complex instructions. The degree of instructions that coders will need usually depends on how socially ambiguous the coded material is. There isn't a lot of ambiguity about whether a movie scene has some food in it or not. There is a great deal of ambiguity about many socially constructed concepts such as "tasty" or "attractive."

Example 11.1E

Table 11.3 shows an example of what the results might look like once the results from Example 11.1D are tabulated.

Table 11.3. Depiction of Occupation and Gender in Women's Magazines, in Percents (*n*)

	Gender		Professional Portrayal	
	Female	Male	At Work	At Play
Percent (***n***)	100 (7)	0	29 (2)	71 (2)

Figure 11.2. Ambiguous cues for gender. Photo by Pimkie, Flickr, licensed under CC BY-SA 2.0.

In the avatar study discussed earlier, the population is, again, role-playing games. The *specific characteristic of interest* is the player's *ability to select for race* when developing their avatar. The research question in this example implies that the coding scheme must include a way for the coder to determine whether or not a gamer has the capacity to choose the racial characteristics of the character when setting up the game conditions.

The coding scheme that the researcher used in the avatar study had three visible markers of race: skin color, hair style and color, and facial features (offered in games that allowed for greater detail). They measured these three markers using the following measures.

Skin tone: The research modified an older scale (the von Luschan scale) used to measure differences in human skin tone. The von Luschan scale distinguishes between 36 different levels of skin tone, ranging from very light to very dark. (The scale was modified to separate tanned from not-tanned skin.) To measure tone, the researcher created the darkest skin color possible to create a character and, using the scale, measured the how much the darkest character possible deviated from white.

Hair color and style: The researcher looked for whether a game allowed players to create a character with an African American hair style, defined as "any hairstyle typically associated with coarse hair (e.g., Afro, dreadlocks, or any hairstyle capable of being worn by someone with coarse hair (no straight hair or loose curls)."

Facial features: The last coding category the researchers used, "African facial features," was based on examining the "size and shape of the nose and lips."

In other words, the researcher was far more specific about measuring race than simply instructing a coder to check for what race the gamer could construct.

Returning to the earlier question of: *Do female characters in strategy video games have unrealistic bodies—overdeveloped breasts and overly narrow waists*? What would be an acceptable coding scheme for measuring breast to waist ratios? Look at the difference between Angelia Jolie's Lara Croft and the character in the original *Tomb Raider* video game. Fairly clearly, using my subjective judgment, Angelina Jolie has a pretty good body shape; the original avatar's body, however, has considerably more breast and far less waist that Jolie. Whatever coding scheme is developed needs to capture that difference clearly.

The researcher could allow the coders to use their own subjective judgment, but we know from earlier research studies that grappled with this question that the "use your own intuition" coding instruction will not work in this case: some coders would code both bodies as unrealistic, some

coders would code both as realistic, and some would code Jolie as realistic and the Lara Croft avatar as unrealistic. In other words, the coders will not agree, and thus *the coding rule is not precise enough that two coders would code the same object the same way most of the time.* However, coders will agree more often using a V body measurement. To measure V, coders should draw an arrow starting at the waist and extend out along the chest wall. The greater the angle of the V, the greater the breast-to-waist ratio is. If the researcher set a rule that "any breast measurement greater than a 50-degree arc is unrealistic," the coders have a precise way to distinguish Jolie (40 degrees) from the Lara Croft avatar (50 degrees).

Figure 11.3 Difference in physical dimensions between a human playing Lara Croft (Jolie in red) versus an early Lara Croft video game avatar (in brown). Illustration by author.

The task for the reader is to judge whether the coding scheme that the researcher has developed is a reasonable way to determine the specific characteristic of study. (The reader has to be able to answer the question, "If a game allows players to create a character with dark skin, an Afro hairstyle, and African facial features, am I willing to agree that the game will allow gamers to construct an African-race avatar? That is, does dark skin, an Afro, and African facial features adequately indicate a particular race? If the V measurement is greater than 50 degrees, does this indicate "unrealistic body measurement"?)

If the answer is "yes," then you—the reader—should accept the coding scheme. If no, then you should reject the coding scheme for this variable and all of the findings based on this coding for this variable.

In other words, judging the adequacy of the coding rule means grappling with how the coding scheme itself defines reality and whether you are willing to accept that reality. Does "unrealistic" mean unreachable by any human, or does it mean unrealistic for most bodies? For example, would you consider Dolly Parton's measurements to be "unrealistic" given that she is a real female who really has a well-endowed figure (and her V measurement is well over 60)? Or would you say that the standard should be the range of measurements for 95 percent of adult women?

12. Judgment Rule 4 for Content Analysis

Judgment Rule: Intercoder reliability should be 80 or above for each variable (some exceptions apply).

Key Takeaways

Judgment rule answers the question: *Is the intercoder reliability high enough?*

Intercoder reliability[1] is a measure of how often two or more coders code the same text (or portion of a text) the same way, it is a measure of consistency. For readers, intercoder reliability is the way to check that the coding scheme (see Chapter 11) is not idiosyncratic or limited to a single individual's opinion, but that coders share a common definition. For example, as discussed earlier, one of the most common coding categories is gender, which for many research projects is either male or female. Look at the picture below. How would you code it?

If you gave this picture to two researchers to code for gender and they both coded this picture as female, then their intercoder reliability score would be 100 percent, or perfect agreement. With 100 percent agreement, then you—the reader—know that whatever is being looked at is fairly clearly defined and understood, at least within the culture that the coders are drawn from (a point to which we will return below). But if two coders do not decide the same way on the same thing, then the reader knows that the coding instructions are not exact enough to be able to tell what the coders are coding. [Excluded content: Figure 12.1 has been omitted from the PDF edition of this text due to licensing limitations. View Figure 12.1 in the chapter online.]

A high intercoder reliability is one of the essential components you need to use to judge the acceptability of the research. As Neuendorf, one of the foremost authorities on content analysis said, "given that a goal in content analysis is to identify and record relatively objective (or at least intersubjective) characteristics of messages, reliability is paramount. Without the establishment of

1. Also called interrater reliability.

reliability, content analysis measures are useless."[2] In fact, "interjudge reliability is [...] the standard measure of research quality. High levels of disagreement among judges suggest weaknesses in [the] research methods."[3]

Intercoder reliability is either measured as percent agreement (from 0 percent, or "no agreement," to 100 percent, or "total agreement"), or by one of several intercoder reliability indexes (which run from zero, "no agreement," to one, "complete agreement"). Each of the intercoder reliability indexes uses a slightly different method to calculate agreement, but in general, most indexes consider .80 agreement as an acceptable level of intercoder reliability.

Intercoder reliability indexes: The most commonly used measures in communications are listed below:

Table 12.1. Chart of Commonly Used Intercoder Reliability Indexes

Index	**Acceptable level**	**NOTE:**
Percent agreement	80% to 100%	Overestimates true agreement
Holsti's method	.80 to 1.00	Overestimates true agreement
Scott's pi (p)	.80 to 1.00	Cohen's kappa is slightly more informative
Cohen's kappa (k)	.75 to 1.00	Considered slightly stricter than other measures
Krippendorff's alpha (a)	.80 to 1.00	Well-regarded
Cronback's alpha	—	Considered inappropriate

Researchers should report what kind of intercoder reliability index they used—Scott's pi, Krippendorff's alpha, or any of the other indexes above—and what level of agreement the coders reached. As a reader, you should look for which index the researcher used and the level of agreement reported. If the researchers report an index with which you aren't familiar, then it is your job as a reader to look up that index and find the level of acceptability for that index.

In truth, there is no completely "set in stone" minimum level of reliability. Most people consider .8 or 80 percent acceptable, and .75 is considered good. *But* for some variables, such as gender, that are considered easy to code, you should probably raise the standard even higher (.93 or better), unless the researcher has given a satisfactory explanation of why a particular population (e.g., babies, the elderly) is particularly difficult to code.

2. Kimberly A. Neuendorf, *The Content Analysis Guidebook* (Thousand Oaks, CA: Sage Publications, 2002), 141.
3. Richard H. Kolbe and Melissa S. Burnett, "Content Analysis Research: An Examination of Applications with Directives for Improving Research Reliability and Objectivity," *Journal of Consumer Research* 18, no. 2 (September 1991): 248, https://doi.org/10.1086/209256.

For hard to code variables, a .75 to .80 intercoder reliability would be acceptable, but lower levels need to be explained. Values as low as .60 to .67 are sometimes reported, but most readers will not accept agreement that low. Remember that if coders randomly guessed gender (without looking at the pictures), they would expect to get the right answer about 50 percent of the time. (A purely random throw of the coin would have a .5 chance of being a head.)

Agreement by consensus. Some researchers use a technique in which they look at all of the disagreements in coding and then "mutually agree" on how to recode that particular disagreement.[4] If the researcher recalculates the reliability index using consensus-derived codes, this is essentially getting two bites at the apple, or research *cheating*. The intercoder reliability *must* be calculated using the original disagreements between coders.

Reporting intercoder reliability. Readers should also look at whether the article writers reported an overall intercoder reliability or reported the reliability on each variable. Ideally, the researcher should report the reliability on each variable. Using the overall intercoder reliability could disguise that some codes are far more unreliable than others. An average of 80 percent on five variables could mean that each variable had 80 percent agreement, or it could mean that four variables had 100 percent agreement and one had 0 percent agreement.

Further Reading

Krippendorff, Klaus. *Content Analysis: An Introduction to its Methodology*. Thousand Oaks, CA: Sage Publications, 2004.

Neuendorf, Kimberly A. *The Content Analysis Guidebook*. Thousand Oaks, CA: Sage Publications, 2002.

4. As Krippendorf points out, the "consensus" or "majority vote" process is deeply flawed: "Observers are known to negotiate and yield to each other in tit-for-tat exchanges, with prestigious group members dominating the outcome. [...] Observing and coding come to reflect the social structure of the group." Klaus Krippendorff, *Content Analysis: An Introduction to its Methodology* (Thousand Oaks, CA: Sage Publications, 2004), 217.

13. Summary of Judgment Rules for Content Analysis

Question

What Is the Research Question?

You need to rewrite the title, or judge from the abstract, or read the article to determine the research question. Once you have the question, you can determine if content analysis is the correct method to use.

If yes, then fine, you can go on to examine more fine-grained aspects of the methodology.

If no, then the details of the method do not matter, because the researcher used the incorrect method to answer their question.

Sample

The Content Analysis Can Only Be Generalized Back to the Research Population Sampled

1. You need to check each time the researchers narrowed the sample. How was the sample selected? How good was the sampling frame? What subgroups would be left out? Would this make a difference?
2. Did the researcher check for external validity? (Did the survey researcher check their sample against the general population?)
    a. If there are no major flaws with selecting the sample, then you can trust that the results for clearly coded variables can be generalized back to the research population.
    b. If the survey sample is *not* representative of the population, then you have a range of answers.

i. The results can be generalized back to the population, with the exception of specific groups that were not sampled. (In other words, the findings probably reflect the population that was actually sampled, and you have a good idea of how these specific sample is different from the general population.)
ii. The results should only be extended to the specified sample. The findings are probably pretty good for the sample, but there are significant unknown differences between sample and the population.
iii. The sampling procedure was so flawed that the findings might actually be misleading. *Do not trust!*

Coding Categories

1. Accept the findings when the coding categories are clear, easy to code, and not leading or biased.
2. Findings (or conclusions) that are based on flawed codes should be thrown out.
3. You can also use the codes to judge a researcher's competence. If there is a consistent pattern of poorly worded or biased codes, you should assume that the researcher was equally sloppy (or biased) elsewhere in the study.

Intercoder Reliability

1. Accept only the variables which have an intercoder reliability in the 75 to 80 percent range (at least). For codes that are extremely straightforward, you should be concerned if the agreement between coders falls below 90 percent.

PART IV

EXPERIMENTAL ANALYSIS

"It doesn't matter how beautiful your theory is [...] If it doesn't agree with experiment, it's wrong."
—Richard P. Feynman

"We must conduct research and then accept the results. If they don't stand up to experimentation, Buddha's own words must be rejected."
—14th Dalai Lama

14. Experimental Analysis Introduction

Routines are comforting and comfortable, and unless something changes, people stick to what has worked in the past. When people deliberately try to change, they usually make the effort because they hope for something better, or because they fear something about what they are doing now. Both the positive (seeking an improvement) and the negative (seeking to avoid) assume some kind of cause and effect between behavior and outcome.

The stronger the link between cause and effect, the more likely that individuals and society will change. Examples from the last few decades include both problems that cigarettes, HIV, and texting while driving cause (lung cancer, AIDS, and car accidents, respectively), and benefits gained from new toys such as iPads and smartphones (the ability to text, the ability to connect to the internet anywhere with tower access, portable game players, music players, and—oh yes—phones). In each case, good or bad, there is a cause and an effect. It is fairly clear (now) that smoking harms the smoker and everyone within breathing distance. The clear and certain link—that smoking *causes* disease[1] and death—has been critical to justifying laws and public policy curbing public smoking. Over the fifty years since the first Surgeon General's report on the dangers of smoking, cigarette use has dropped by more than one half.[2] Societal pressures have driven smoking out of college campuses, airports, restaurants, bars, and many other public spaces.[3]

1. The Surgeon General's report on fifty years of progress in reducing smoking says, "This report finds that active smoking is now causally associated with age-related macular degeneration, diabetes, colorectal cancer, liver cancer, adverse health outcomes in cancer patients and survivors, tuberculosis, erectile dysfunction, orofacial clefts in infants, ectopic pregnancy, rheumatoid arthritis, inflammation, and impaired immune function. In addition, exposure to secondhand smoke has now been causally associated with an increased risk for stroke." Preface by Boris D. Lushniak, Read Admiral, U.S. Public Health Service, from the U.S. Department of Health and Human Services, *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General* (Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2014), iii, https://www.ncbi.nlm.nih.gov/books/NBK179276/.
2. Adult smoking has decreased from 42 percent of the population in 1965 to 18 percent in 2012.
3. Kathleen Sebelius, "Message from Kathleen Sebelius," from the U.S. Department of Health and Human Services, *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General* (Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2014), https://www.ncbi.nlm.nih.gov/books/NBK179276/.

15. Judgment Rule 1 for Experimental Analysis

Judgment Rule: Experiments should be used to determine causality.

Key Takeaways

Judgment rule answers the question: *Is the researcher interested in determining causality?*

Using Experiments to Determine Causality

Once again, the first question to ask when looking a research paper is: What question is the researcher trying to answer, and what method is appropriate to answer this question?

If the researcher is interested in finding out the characteristics of a population—for example, how many people own and use smart phones—then a survey is the best approach. If the researcher is interested in finding out if using smartphones destroys people's ability to cross the street safely, then the researcher is asking a causal question: "Do smart phones decrease pedestrian safety?"

Scientists use the experimental method to determine if A *causes* B. The core of the experimental method is to control all possible explanations of why the experimental group changes. If the scientist can create a situation where *just one thing is changed* (A), and that one change also *changes something else* (*produces* B), then we can fairly safely say that A *caused* B.

Each question in Example Box 15.1 is asking a causality question, which you can also think of as a question about sequence. In an experiment, as opposed to a survey, you know what happened first and what happened second. Pretty obviously, if watching television commercials for high-density, high-calorie food *causes* preteens to prefer high-density, high-calorie food, then the commercials should come first, and the increase in wanting high-density, high-calorie food second.

Example 15.1

Research questions that an experiment can answer

Does watching television commercials for high-density, high-calorie food increase preferences for high-density, high-calorie food in preteens?

Does watching news online decrease comprehension of issue-based (thematic) news stories?

Does using highly sexualized female avatars decrease playing competence in first-person shooting games?

Does using Facebook increase motivated performance on exams?

Does using Twitter increase the subject's ability to think logically?

Let's go into the difference between survey and experiment in a bit more detail.

Let's say that a researcher surveyed a group of teenagers and college students about their video-gaming habits. The researcher asked students what games they play, what avatars they used, what those avatars looked like, and what their average scores were (or the highest levels they reached) on each of their games. The researcher then compared the average scores of the group of players who used highly sexualized avatars against the group of players who did *not* use highly sexualized avatars and found out that those people who used highly sexualized avatars did *not* score as highly (or reach as high levels) as those without sexualized avatars.

What has this researcher found out from this survey? Well, as it turns out, the researcher now knows exactly what was said above—that people who used highly sexualized avatars did *not* score as highly (or reach as high levels) as those without sexualized avatars. In other words, the researcher knows that sexualized avatars and low scores co-vary. (The term "co-vary" essentially means that when one thing changes, another thing changes in a systematic way.) What the researcher *doesn't* know is what caused the change. Did the people who used the highly sexualized avatars do less well because they saw themselves as less capable, or did more skilled, more dedicated players choose less sexualized avatars because they were more interested in the gaming (the shoots, the kills, achieving levels) than in the presentation of the avatar? What came first? What influenced what?

In a well-designed experiment, the researchers know that the treatment alone caused the change. That is, the scientists have designed their experiment in such a way that they can rule out all other

potential causes. Let's take the simple research question: Does listening to rap music in the car slow braking time? To find out the answer, researchers gathered a bunch of students who were practicing for their first driving license, gave them driving tests in a virtual driving simulator, and found out that the students who listened to rap music *improved* their braking time.

To check whether the experimental treatment (in this case, rap music) caused the change, researchers *controlled* the experiment. One straightforward way to control whether the treatment or something else causes a change is to give the experimental treatment to half of the research subjects (the experimental group), and to not give the treatment to the other half (the control group) (See Figure 15.1). If the two groups start out as similar and both go through the same processes during the experiment, then the differences between the experimental condition and the control condition will be due to the treatment, and only the treatment.

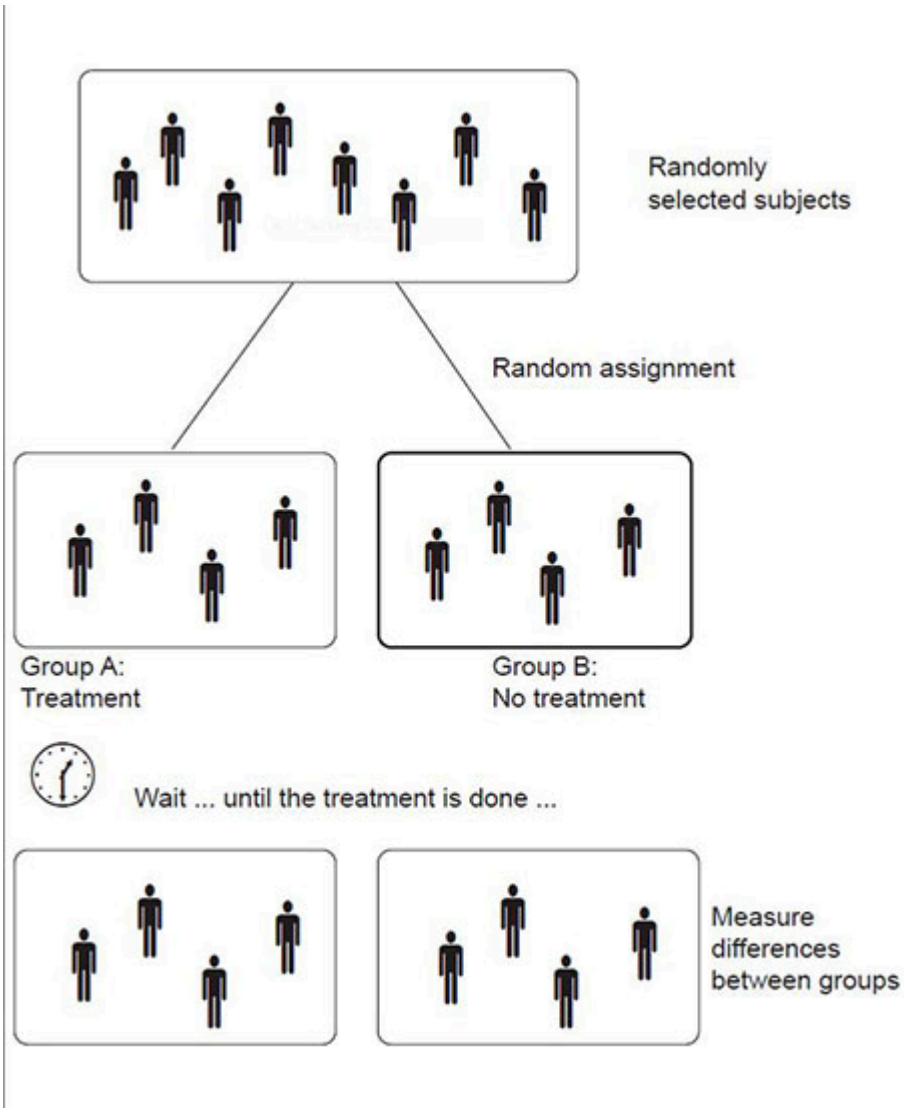


Figure 15.1. Diagram of the sequential flow of the experimental research process

Even before an experiment is run, the researchers should be able to describe what results they could get and what each of those results would mean in terms of answering their original question: Does listening to rap music increase braking time?

> Potential result 1: If both the experimental and control groups improved braking speed, then "something else" caused the decrease in braking time (not the rap music). (The students probably learned to be better drivers, but this is a guess because the researchers didn't control for learning how to drive.)
>
> Potential result 2: If neither group improved, then the music had no effect.
>
> Potential result 3: If the control group learned how to brake more effectively than the experimental group, then the treatment *harmed* the driver's ability to brake, *even if* the experimental group's braking scores improved over the course of the experiment.
>
> Potential result 4: If the control group braked less effectively than the experimental group, then the rap music improved the driver's ability to brake, *even if* the control group learned how to brake more effectively during the experiment.

Summary: Controlled experiments are, by definition, set up so that researchers can manipulate "A" to see if "B" changed. How the researchers run the experiment is called subject design, and is the next basic judgment you will need to make. Is the experiment designed appropriately?

16. Judgment Rule 2 for Experimental Analysis

Judgment Rule: The experimental design should rule out other potential explanations for any changes observed during the experiment.

Key Takeaways

Judgment rule answers the question: *Did the researcher use an appropriate experimental design?*

The two most basic experimental designs are: (1) determining if an experimental treatment will produce observable differences between two otherwise identical groups (*between group design*), or (2) determining if a treatment changes an individual from time 1 to time 2 (*within-subject design*, also known as *repeated measurement design*).

Between Group Design

For between group designs, researchers equalize the experimental and control groups by randomly assigning subjects to groups. As a reader, look for whether the subjects were randomly assigned to groups. (Typically, research papers have the following sections: Introduction, Review of Literature, Methods, Findings, and Discussion. Information on how the subjects were assigned to experimental and control groups is in the methods section of the research paper.)

Theoretically, randomization should adequately even out all differences between groups, but researchers are often interested in understanding specific characteristics of the population rather than just randomly assigning these differences away. For example, going back to the research questions discussed earlier, each different question in Example 16.1 has subgroups about which we might specifically want more information. For example, we know from a lot of research that people with a poor body image handle food pressures differently than people with a good body image, so

we might reasonably expect kids with a poor body image to be more affected by television commercials for high-density food (See Example 16.1 for additional subgroups).

Example 16.1

Question 1: Does watching television commercials for high-density, high-calorie food increase preferences for high-density high calorie food in pre-teens?

Potentially important subgroups:

Subgroup condition 1: Obese subjects versus normal weight subjects versus underweight subjects.

Subgroup condition 2: Low body image versus normal body image.

Question 2: Does using highly sexualized female avatars decrease playing competence in first-person shooting games?

Potentially important subgroups:

Subgroup condition 1: Males versus females.

Subgroup condition 2: Self-objectifying versus non-objectifying subjects.

Question 3: Does using Facebook increase motivated performance on exams?

Potentially important subgroups:

Subgroup condition: Self-presenters versus self-disguisers.

And, of course, kids with a poor body image are already at risk for eating disorders, so there might be important policy implications in finding out if this specific group reacts differently. We also know, from previous research, that girls who objectify their own bodies are less likely to perceive themselves as competent at various physical and cognitive skills. Would they also be more affected by using a highly sexualized avatar? And finally, we know from recent research that people have different ways of presenting themselves on Facebook. Some people present a highly positive (even self-congratulatory) self-image—one that consistently shows themselves at parties and fun places, doing fun things with good-looking, smiling people. These people tend not to self-disclose negative information, which means that their Facebook "support" for mutual disclosure (and emotional support) is fairly weak. Others choose a much more open strategy in which they balance the good with the bad, and these Facebook presenters are much more likely to receive emotional support from their Facebook friends when they need it. Given that students might need some emotional support during exams, the researchers reasoned that presentation style might affect exam per-

formance, leading to the research question "Does the Facebook strategy that people use make a difference in determining motivated performance on exams?"

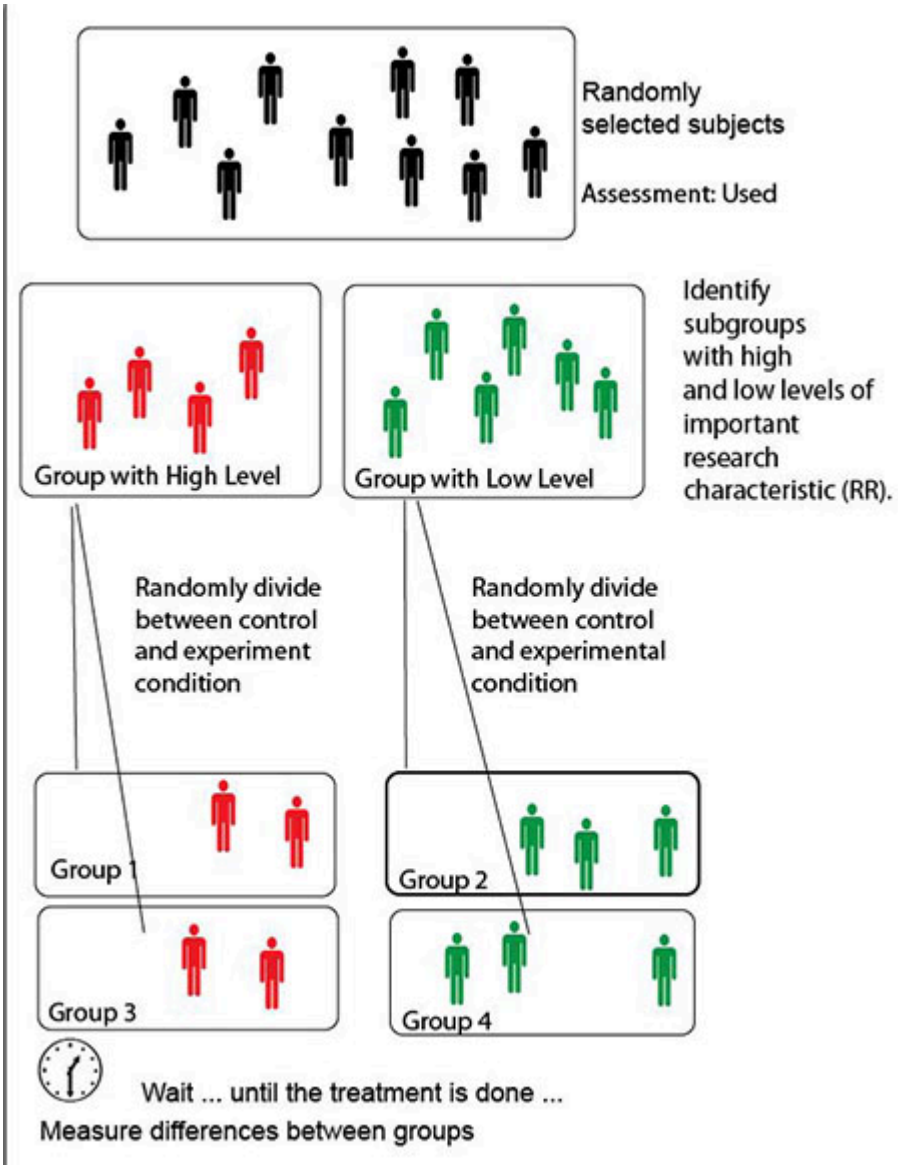


Figure 16.1. Experimental method, dividing by condition

How, then, best to determine if specific subgroups respond to the experimental treatment differently than the group as a whole? To figure out treatment effects on these subgroups, researchers use an additional step before they divide the subjects into control and experimental groups (see Figure 16.1). In this step, the researchers give all subjects a preliminary test for the characteristic of interest—body image for the research question on preferences for high-density, high-calorie food; self-sexualization for the research question on the effects of using highly sexualized images; and self-presentation strategies for the question on motivated performance on exams.

The researcher then divides the original subjects into two groups based on the results of the assessment, and then randomly divides each group into a control condition and an experimental condition.

Turning to a specific example, a researcher who is interested in testing whether commercials for high-density and high-calorie food trigger desire for high-calorie food in pre-teens would first gather a group of preteens and give them a test for body image.[1] After the children took the test, the researchers would be able to distinguish children who had a good body image from those who were dissatisfied with their bodies. Next, the researcher randomly distributes the normal body image group into a normal body image control group and a normal body image experimental group, and the dissatisfied body image group into a dissatisfied control group and a dissatisfied experimental group, so the researcher will be able to compare the changes from all conditions with each other. As you would expect, the researchers will be able to tell if the treatment had an effect by comparing the results of both experimental groups with both control groups. They will also be able to tell if the low body image group subjects were systematically more (or less) influenced by the television commercials than the normal body image group.

Within-Subject Design

The within-subject design is the other major type of experimental design. In this design, the experimenter compares the scores of a single individual over time. Each participant becomes his or her own control, and all participants are exposed to every treatment. In within subject designs, researchers are interested in the change in each subject between time zero (before the treatment starts) and subsequent times (after the subject has received a treatment).

Within-subject designs are considered quite sensitive because the researcher is tracking the change in a single individual.

1. An example of this type of scale is the Children's Body Image Scale. Helen Truby and Susan J. Paxton, "The Children's Body Image Scale: Reliability and Use with International Standards for Body Mass Index," *British Journal of Clinical Psychology* 47, no. 1 (March 2008): 119-124, https://doi.org/10.1348/014466507X251261. For this test, children are given seven photographs of children whose weight ranges from very thin to obese. The children are told to select which picture they believe looks most like their size (perceived weight) and which picture represents the size they would like to be (ideal weight). The measure of body dissatisfaction was the difference between the size the child said he or she would like to be and the size the child thought he or she was.

> An advantage of within-subjects design is that individual differences in subjects' overall levels of performance are controlled. This is important because subjects invariably will differ greatly from one another. In an experiment on problem solving, some subjects will be better than others regardless of the condition they are in. Similarly, in a study of blood pressure some subjects will have higher blood pressure than others regardless of the condition. Within-subjects designs control these individual differences by comparing the scores of a subject in one condition to the scores of the same subject in other conditions. In this sense each subject serves as his or her own control.
>
> David M. Lane, "Experimental Designs," *Online Statistics Education: A Multimedia Course of Study* (Rice University, University of Houston Clear Lake, and Tufts University, last accessed June 2, 2023), http://onlinestatbook.com/2/research_design/designs.html.

But again, the researcher—and the reader—need to consider whether the treatment made the difference or whether an observed change was due to something else that happened during the experiment. Did the subject's score decrease because of the experiment or because he or she got tired? Or bored? Did the subject's score get better because he or she learned how to perform a task more effectively? Changes in performance due to fatigue, boredom, or learning would not be because of the *treatment*, but because *the test subject changed for some other reason.*

To control for these test differences, researchers generally counterbalance (switch) the order of the treatments. So in an experiment with two treatments, half of the subject would receive treatment 1 followed by treatment 2, and the other half would receive the treatment 2 first and treatment 1 second. If there were no differences due to the order of the test, then any differences observed would be from the experimental treatments.

Let's say that a researcher is interested in comparing whether violent movies increase viewer aggression more than or less than violent video games. In a within-subjects design, the subjects would be shown violent movies, tested for aggression, and then allowed to play violent video games, followed by a test for aggression. It is possible that the first treatment would increase each subject's overall aggression level, thus making it more likely that the second treatment would show an aggressive response (as suggested by excitation transfer). Researchers check for this carry-over effect by randomly assigning subjects to alternating treatments. Half of the subjects would watch the movie first and then play the video game, while the other half would play the video game first, followed by the movie. If the violent video games increased aggression, then the subjects who played the game should test as having higher aggression scores after playing the game regardless of whether they played the game first or second.

Alternating treatments in this way is called counterbalancing. The aggression scores from the group watching a movie first will be counterbalanced with the other group (those who played violent video games first).

Judgment Rule 2B

Key Takeaways

Judgment Rule 2B: *Did the researcher randomly assign subjects to groups (between subjects) or to treatment order (within subjects)?*

Randomization and Experimental Design

Randomization in experimental methods is, and is not, like randomization in surveys. For surveys, researchers are concerned with whether bias has been introduced in the selection of the sample from the overall population. However, the bias that experimenters (and readers) are concerned about in experiments is whether there is a difference between the experimental and the control conditions (between groups), or between counterbalanced groups (within subject conditions). That is, the experimenters are concerned with whether they introduced bias when assigning subjects to groups; they are typically not concerned with whether the original experimental sample is like a larger population. In a survey, for example, researchers would care if everyone in their sample were left-handed, Welsh vegans (who were either highly aggressive or not aggressive at all). An experimenter wouldn't care as long as all groups had equal numbers of highly aggressive subjects.

For both surveys and experiments, the randomization process reduces the change of bias; it doesn't eliminate the possibility. You could randomly pick ten pieces of your favorite dark chocolate out of a bag of mixed chocolates, but it is highly unlikely. More likely, you picked out the dark chocolate you liked over the milk chocolate you didn't. Randomization guards against bias in the selection process.

While it is the responsibility of the researcher to take steps to guard against introducing bias, it is the reader's responsibility to check whether the researcher took those steps (for example, assigned subjects randomly).

17. Judgment Rule 3 for Experimental Analysis

Judgment Rule: Accept the findings only if and when the treatment could reliably be expected to produce the change for which the experimenter is testing, and when the control would reasonably be expected to not produce this effect.

Key Takeaways

Judgment rule answers the question: *What is (potentially) causing a change in the experimental participants?*

The treatment is the very specific way that the researcher manipulates the experimental subjects (see Example 17.1). The research question sets up what potential "cause" is being examined. Usually the cause is a category—high-fat, high-calorie food (Question 1); online news (Question 2); highly sexualized female avatars (Question 3). But researchers have to test something specific. In other words, they cannot just test a generic high-fat, high-calorie food; they must use a lasagna, or a burger, or a cake. The experimental treatment is the specific food, advertisement, movie, picture, news story, or avatar used. It is part of the researcher's job to write a good enough description of the treatment that you, the reader, can visualize exactly how the researchers manipulated their subjects, and it is part of the reader's job to think about—and judge—how well the specific treatment represents the general category that the researcher is looking at. *Doom* or *Call of Duty* would be appropriate treatment games for experiments that were looking at the impact of violent shooter games. Tetris, a tile-matching puzzle game, would not.

The reader, then, needs to first look at the research question to figure out what the researcher is testing. In the first research question in Example 17.1, the researchers are looking at the impact of high-density, high-calorie food. Therefore, any commercial for a specific food that had the qualities of high-density, high-calorie food would qualify as a reasonable experimental treatment—a Domino's thin crust, 14-inch pizza with extra cheese (250 calories a slice, or 2000 for an entire

pizza);[1] a Hardees half-pound Texas BBQ Thickburger (1030 calories)[2]; or a Culver's chocolate, two-scoop waffle cone (657 calories).[3]

Example 17.1

Research Questions an Experiment can Answer

Research question: Does watching television commercials for high-density, high-calorie food increase preferences for high-density, high-calorie food in preteens?

Experimental treatment: Television commercials that feature high-density, high-calorie food. (Any of the following commercials could be used as examples of the kind of food needed for the experimental manipulation: a Domino's pizza, a Hardees half-pound Texas BBQ Thickburger, or a Culver's double waffle cone.)

Research question: Does watching news online decrease comprehension of issue-based (thematic) news stories?

Experimental treatment: In this experiment, the researchers would need to test reading comprehension of the same story online (experimental group) and in print (control group). Since the research question specifically called for the more complex stories that deal with the causes and the implications of social issues (thematic stories), the researchers would also need to select some longer, in-depth news stories for both groups to read, such as: What are the root causes of the Gaza occupation? What are the implications of keeping live samples of the smallpox vaccine in research facilities?

Research question: Does using highly sexualized female avatars decrease playing competence in first-person shooting games?

Experimental treatment: Having some video players use a highly sexualized avatar (experimental group), and some use a non-sexualized avatar (control group).

Research question: Does using Twitter increase subject's ability to think logically?

1. Calories are for a 14-inch pizza with extra cheese and thin crust. "Cal-O-Meter," Domino's Pizza, last accessed June 2, 2023, https://www.dominos.com/en/pages/content/nutritional/cal-o-meter.
2. Calories for a half-pound Thickburger. "New Texas BBQ Thickburger from Hardee's and Carl's Jr.," Grub Grade, archived January 27, 2015, https://web.archive.org/web/20150127080436/http://www.grubgrade.com/2014/07/21/new-texas-bbq-thickburger-from-hardees-and-carls-jr/.
3. "Culver's Chocolate Waffle Cone (2 Scoop)," Fatsecret, last updated July 10, 2022, https://www.fatsecret.com/calories-nutrition/culvers/chocolate-waffle-cone-%282-scoop%29.

Experimental treatment: Sending tweets in Twitter format (experimental group) versus having unlimited length in text message (control group).

The experimental treatment (for the experimental group) also needs to be paired with a control. The control mimics the same characteristics of the treatment *without* the specific aspect of the treatment that the researcher is interested in studying. The classic example is the sugar pill (placebo). Researchers have found out that just the act of swallowing a pill can improve people's satisfaction with their health—in part because people believe that medicine will work, even when what they are actually given is a lump of sugar.[4] The control "treatment" runs a group through a similar process as the experimental treatment, without the specific features that the experiment is testing for. Going back to the high-fat, high-density food example, a number of different commercials could be used as controls—commercials for cars, perfume, cranberry, apple, or travel. The researcher should be careful to match as many characteristics of the two sets of commercials as possible—length, production quality, known brands—but the experimental treatment must have the distinguishing research characteristic, and the control cannot have this characteristic.

4. The placebo effect is powerful. When we think that something will work, a drug is, in fact much more likely to be effective. In act some searchers have suggested that a "substantial percent of the effects from antidepressants may be placebo effect." For a quick discussion of the placebo effect, see Christopher Lane, "Placebos Do Work," *Psychology Today*, June 26, 2009, http://www.psychologytoday.com/blog/side-effects/200906/placebos-do-work-lets-consider-why.

18. Judgment Rule 4 for Experimental Analysis

Judgment Rule: Accept the findings only if the dependent variable can reasonably be expected to detect changes in the experimental participant.

Key Takeaways

Judgment rule answers the question: *Can the dependent variable reasonably be expected to detect any changes the treatment may produce?*

Obviously, the treatment phase of the experiment—if the treatment actually is effective—produces a change in the treated subjects. The next phase of the experiment involves measuring what (and how much) change was produced. Researchers call this the testing phase, and that thing that measures change is called the dependent variable.

Researchers generally have a variety of measures that they could use to as the dependent variable (see Example 18.1). For example, researchers could:

> Ask subjects to rank their preferences for food on a list that has some high-fat/high-density items and some low-fat/low-density items,
>
> Ask subjects to list what food they intend to eat at lunch, or
>
> Give subjects a buffet and see what they eat (revealed preference).

All these measures can detect changes in people's food preferences and so meet the basic criteria for an adequate dependent variable. Some are actually stronger measures than others—for example, watching what kinds of foods that people eat is a stronger measure of food preference than asking people what food they intend to eat at lunch (revealed preference is a stronger measure than stated preference), but all measures will give the researcher some indication of whether preteens exposed to high-calorie, high-density food commercials changed their food preferences.

Example 18.1

Examples of Dependent Variables

- Researchers want to know if listening to music can lower blood pressure. In this experiment, blood pressure measurements are the dependent variable.
- Researchers want to know if Facebook activity reduces prosocial behavior. In this experiment, willingness to give money to strangers in a dictator game was the measure of prosocial behavior.
- Researchers want to know if reading, watching, or listening to news increases depression. In this case, the Beck Depression Inventory was the dependent variable.

The basic judgment rule for this phase of an experiment is: Can the dependent variable reasonably be expected to detect the changes that the treatment produced? In forming a judgment about the quality of the dependent variable, the reader needs to consider two basic questions: First, is the dependent variable reliable? Second, is the dependent variable valid?

Reliability: Reliability is defined as the degree to which the instrument used to quantify the dependent variable gives the same reading each (and every) time. If a researcher wanted to measure the relative importance of race in a local newspaper, one simple way to measure "importance" is to physically measure the number of column inches of newsprint. A researcher who used a steel ruler, which is a relatively inflexible instrument, will probably get the same number of inches of newsprint—or something very close—each time he or she measured a column of print. If researchers used something more flexible—a rubber band marked in inches, or silly putty—they might reasonably expect to get different answers each time they measured the column. The silly putty scale would be considered unreliable.

Researchers have several checks for reliability. Some of the major checks are:

Inter-rater reliability: Two researchers (coders) get the same answer using the same measure.

Checking for intercoder reliability:

As a reader, you need to look for whether the researcher measured how consistently coders agreed with other coders' ratings. Perfect inter-coder reliability means that every coder rated every coded variable exactly the same way (a reliability score of 1). No correlation between coders would be a reliability score = 0: the coders disagreed on every measure. (Notice that you will judge

the reliability of coding for experiments exactly the same way as judging intercoder reliability for content analysis.) Table 18.1 lists the generally accepted guidelines for judging intercoder reliability:

Table 18.1. Guidelines for Acceptability of Intercoder Reliability

Guidelines for Acceptability of Intercoder Reliability	
.9 and greater	Excellent reliability
.8 to .9	Good reliability
.7 to .8	Acceptable reliability
.6 to .7	Questionable reliability
.5 to .6	Poor reliability
Under .5	Unacceptable reliability

Test-retest reliability: An assessment measure produces the same answer time and time again. (A thermometer that gives you a 105-degree fever three times in fifteen minutes has high reliability—and also suggests that you should go to the emergency room.) If you took your temperature three times in quick succession and got respective readings of 86 degrees, 106 degrees, and 93 degrees, then your thermometer isn't reliable.

Test-retest reliability: Test-retest reliability is determined by repeatedly measuring the same respondents (sometimes with the same question a few minutes apart, say at different points in the survey, or by testing subjects twice). Like inter-coder reliability, readers need to look at whether the researcher reported test-retest reliability. Test-retest reliability is measured with a correlation coefficient which can range from 0 (perfect unreliability) to 1 (perfect reliability). Test-retest reliability of over .7 is generally considered acceptable.

Internal consistency reliability: Researchers ask multiple questions on the same construct and check whether the subjects answer the questions consistently. The Beck Depression Inventory (see Example 18.2) is a pen-and-paper test used to detect depression. The basic construct is "depression." The inventory has twenty-one questions, each one of which asks a about a different aspect of level of unhappiness. The idea behind internal consistency is that if the questions all measure the same latent construct, then a subject's answers should all generally agree with each other.

Example 18.2

Beck Depression Inventory

For all questions, please answer with regard to the last two weeks.

1. How is your mood?

    ___ I do not feel sad

    ___ I feel blue or sad

    ___ I am blue or sad all the time and I can't snap out of it

    ___ I am so sad or unhappy that I can't stand it

2. How pessimistic are you?

    ___ I am not particularly pessimistic or discouraged about the future

    ___ I feel discouraged about the future

    ___ I feel I have nothing to look forward to and I won't ever get over my troubles

    ___ I feel that the future is hopeless and that things cannot improve

3. Do you feel like a failure?

    ___ I do not feel like a failure

    ___ I feel I have failed more than the average person

    ___ As I look back on my life all I see is a lot of failures

    ___ I feel I am a complete failure as a person

4. Are you satisfied?

    ___ I do not feel particularly dissatisfied

    ___ I feel bored most of the time and don't enjoy things I used to

    ___ I don't get satisfaction out of anything anymore

    ___ I am dissatisfied with everything

5. Do you feel guilty?

    ___ I don't feel particularly guilty

    ___ I feel bad or unworthy a lot of the time

___ I feel quite guilty and bad or unworthy practically all the time

___ I feel as though I am very bad or worthless

Internal consistency: Looking at internal consistency is important when the researcher's dependent variable is a scale developed from a list of questions about the same construct, as in a list of questions that together measure "body image," or "acceptance of rape myths," or "depression." To determine whether the researcher has checked for internal consistency, you check whether the items on a scale vary together. For example, you would logically expect that a person who says, "I am so unhappy that I cannot stand it" (see Figure 18.2) is more likely to say, "I feel that the future is hopeless and that things cannot improve" than a person who says either, "I do not feel sad" or, "I am not particularly pessimistic or discouraged about the future." And, in fact, research studies over decades and from many countries have found that people are consistent in how they respond to this test.

Researchers have two primary ways to report internal consistency. First, the researchers can measure the consistency of a set of items themselves. Internal consistency is most commonly measured by Cronback's alpha, with under .5 as unacceptable reliability and above .7 as acceptable.

Second, the researcher can use an established scale—such as the Beck Depression Inventory—where other researchers have measured and established reliability. To check, look for whether the researchers in the article you read referred to a study that established a scale or a test's reliability.

Judgment Rules for Reliability

The reader's check for whether the dependent variable is reliable depends on the exact method used to check reliability. For both intercoder reliability and test-retest reliability and internal consistency tests, look for whether the researchers' measure of reliability is over .7. As usual, if the researcher does not report this measure, you should assume that the test is *not* reliable.

Validity refers to the degree to which the instrument measures what it is supposed to measure. Remember, reliability is defined as being able to measure the same thing in the same way. It is possible to construct a highly reliable measure that is also totally wrong. Let's say, for example, that you were constructing a test for male babies by clothing color, and you developed a measure that coded babies wearing pink as males and wearing blue as females. It is highly likely that you could get a reliable measure. Coders are likely to be able to reliably and consistently tell what babies are

wearing pink and what babies are wearing blue. You would also most likely have a terribly invalid measure, particularly in this culture, where male babies are traditionally dressed in blue and girl babies in pink.

Once again, "valid" means that researchers are studying what they think they are studying. There are two basic ways readers should judge validity: face validity and construct validity. But before discussing these, let's turn to one measure of validity that is *not* acceptable, faith validity.

Example 18.3

Partial list of questions used to test for rape myth acceptance. Source: Diana L. Payne, Kimberly A. Lonsway, and Louise F. Fitzgerald, "Rape Myth Acceptance: Exploration of Its Structure and Its Measurement Using the *Illinois Rape Myth Acceptance Scale*," Journal of Research in Personality 33, no. 1 (March1999): 27-68, https://doi.org/10.1006/jrpe.1998.2238; Sarah McMahon and G. Lawrence Farmer, "An Updated Measure for Assessing Subtle Rape Myths," *Social Work Research* 35, no. 2 (June 2011): 71-81, https://doi.org/10.1093/swr/35.2.71.

Honesty Subscale for Rape Myth Acceptance Scale	1	2	3	4	5
Subscale 1: She asked for it					
1. If a girl is raped while she is drunk, she is at least somewhat responsible for letting things get out of hand.					
2. When girls go to parties wearing slutty clothes, they are asking for trouble.					
Subscale 2: He did not mean to					
1. Rape happens when a guy's sex drive goes out of control.					
2. It shouldn't be considered rape if a guy is drunk and didn't realize what he was doing.					
Subscale 3: It wasn't rape					
1. If a girl doesn't physically resist sex—even if protesting verbally—it cannot be considered rape.					
2. If a girl doesn't say "no," she can't claim rape.					
Subscale 4: She lied					
1. A lot of times, girls who say they were raped agreed to have sex and then regretted it.					
2. A lot of times, girls who claim they were raped have emotional problems.					
Scoring: Scores range from 1 (strongly agree) to 5 (strongly disagree). Scores may be totaled for a cumulative score. A higher score indicates a greater rejection of rape myths.					

Faith validity: Faith validity is simply blind faith that a measure works. Without empirical evidence, without testing, the researcher claims a test is valid because the researcher believes the test is valid. Faith validity is particularly problematic because the researcher's faith in the measure can

also draw the reader into accepting the researcher's biases. Just because the researcher labels a scale as "Honesty" does not—in and of itself—mean that the scale can measure honesty. Take, for example, Example 18.3. Do these items really measure how honest a person is? No, not really. All of the questions are really about what situations qualify as rape and what situations don't qualify as rape. The questions do not test attitudes about honesty, nor can they distinguish when the person taking the test is lying.

On the face of it, the questions on the honesty scale are not a valid test for honesty just because the scale is labeled "honesty." (The scale is actually a test for rape myth acceptance.)

Readers need, instead, to rely on their personal assessment of a dependent measure's validity (face validity), or to look for how the researchers tested for construct or empirical validity (construct validity).

Face validity: Face validity is a straightforward judgment of whether the questions are reasonably able to measure what they are supposed to measure. For example, going back to the Beck Depression Inventory, it seems reasonable that someone who says, "I am so unhappy that I cannot stand it," and, "I feel that the future is hopeless and that things cannot improve," is—at least for the moment—depressed, while people who say that they "do not feel sad at all" and that they "are not particularly pessimistic or hopeless about the future" are not likely to be depressed. So, a "best guess" assessment is that these questions are likely to show which subjects are depressed and which subjects are not. The most important word in the previous sentence is "likely." Face validity means that the test looks like it will work, and will probably work, but there is no real testing to determine if the test actually works.

Construct validity: Researchers have a variety of methods they use to test for construct validity. One, researchers use a panel of "experts" to judge whether the test questions tap into the different aspects of the main construct. Two, the researcher (or reader) considers whether the test includes questions about all aspects of a construct that the theory suggests are important. Anxiety, for example, alters behavior. (Anxious people tend to startle more easily, have more trouble sleeping through the night, and are less able to sit still.) Anxiety also changes people's judgment. Anxious people are more likely to predict disaster, to catastrophize, and to worry. And, of course, anxiety is an emotional state. A researcher who is testing whether horror movies increase anxiety should use a scale that tests for all of the different aspects of anxiety: judgment, behavior, and emotion.

Empirical validity/concurrent validity: Empirical validity tests how closely scores on a test correspond to some other measure that has already been established. For example, how well does the Beck Depression Inventory test for depression? Would the Beck Depression Inventory and a group of highly skilled psychologists identify the same patients as depressed? The test of empirical validity would be the degree to which both the Beck Depression Inventory and the group of psychologists agreed with each other. A valid test would show that the patients the psychologists

identified as depressed people gathered on the high end of the Beck Depression Inventory scale, and the not-depressed people gathered on the other end of the scale. (They do.)

Predictive validity: Another method of looking at validity is predictive validity—determining whether the test can predict future behavior.

Judgment Rules for Validity

In thinking about whether to accept the researcher's assurances about validity, readers have two checks. First, do you personally think that the checks that the researcher used to detect treatment effects are reasonable? If so, why? For example, in one study of altruism, the researcher tracked whether students leaving the experimental room held the door open for a research assistant (who was, for the experiment, pretending to be on crutches). Do you agree that holding a door open for a person using crutches is a helpful act? What about the opposite? Is a person who didn't hold the door open not helpful?

If you have two groups, one that saw an action film and another that saw a chick flick, and the group that saw the chick flick was far more likely to open doors than the group that saw the action film, would the experiment show an increase in helpfulness in the chick flick film people? Or did it show a decrease in helpfulness in the action film group? Either way, is it safe to say that helpfulness was affected? If you—the reader—think that opening doors is a valid measure of helpfulness, then—yes—you would have to say that the experiment showed that the genre of films seen impacted the subject's willingness to offer aid and comfort to the poor research assistant on crutches.

Second, you should look for information about how much the measure used has been tested for validity. Some measures, like the Beck Depression Inventory, have been used to test for depression for decades, and most researchers assume that readers will know and accept the validity of this test. But readers who are learning the field or readers who encounter an unfamiliar test will need to do some extra work to check out the measure's validity. To illustrate, a search of the terms "Beck Depression Inventory," "reliability," and "validity" will turn up scores of articles that measured the reliability and validity of this test. For the less well-known tests, authors will commonly cite previous studies that have validated the measures that they use. As a reader, you should look for these citations and, if you have any concerns about the measure or if the researcher did not report the questions used to develop the measure, *go to the original article*, and look at how the measure was developed.

19. Summary Judgment Rules for Experiments

Research Design and Sampling Method

The primary purpose of an experiment is to test for causality, not to describe a population. To determine causality, you need to be able to compare the difference between two groups—one treated and one controlled– or the change in one person from time A to time B.

Research Design

If you are looking *between groups*: look for whether the control and the experiment group were identical before the experimental treatment started. If there are no systematic differences between the control group and the experimental group, then you can trust that differences detected are due to the treatment. For between-group designs, researchers equalize the experimental and control groups by randomly assigning subjects to the groups. *Look for whether the subjects were randomly assigned to groups.*

> *If differences between the subjects in the control and experimental groups could also produce the same impact that the treatment is trying to produce, then the experiment is fatally flawed.* DO NOT TRUST!

OR

> For *within subject* designs, all participants are exposed to every treatment (or condition). Each participant becomes his or her own control. The researchers, however, need to check whether the *order* of the testing has made a difference— that is, whether something (other than the treatment) has caused a change during the experiment.
>
> The reader should *check whether the researcher has counterbalanced the treatment order and randomly assigned subjects to the different treatment sequences.* IF NOT, *DO NOT TRUST.*

Treatment

Accept the findings only if and when the treatment could reliably be expected to produce the change for which the experimenter is testing, and *when the control would reasonably be expected to not produce this effect.* IF NOT, DO NOT TRUST.

Dependent Variable

Accept the findings only if the dependent variable can reasonably be expected to detect impact. IF NOT, DO NOT TRUST.

If all considerations for trust are met, then trust the experimental results.

PART V

SUMMARY AND CONCLUSIONS

"Life is not found in atoms or molecules or genes as such, but in organization; not in symbiosis but in synthesis."

—Edwin Conklin

Evolution by Association : A History of Symbiosis: A History of Symbiosis, Oxford University Press, 1994.

"[S]cience consists of theories or insights arrived at as a result of systematic reflection or reasoning. ... It involves, therefore, the analysis of experience and the synthesis of the results of analysis into a comprehensive or unitary conception."

—Joseph Alexander Leighton

The Field of Philosophy, Wentworth Press, 2019.

20. Compiling a Summary of Research Findings

The goal of research is to build a picture of the world.

The picture that develops will depend on the question asked.[1] Is it a research question? A policy question? A moral question? An engineering question? Something else?

A research question essentially asks, "What is?" What do we know about hypermasculinity? How much violence is there in media? What impact does priming have on the brain? Does having smartphones out during conversations decrease people's satisfaction with the quality of the interaction? All these questions ask for descriptions of a) what is knowable (verifiable), b) what is known, and c) how it is known (assessment of the method's soundness).

Moral questions address the issue of "what should be," and only need research to verify statements of fact. For example, let's examine what research would be useful for answering the question, "Is it wrong to have a television show that portrays Lucifer as an appealing character?" Fox's announcement of *Lucifer*, a television series based on the premise that Lucifer is bored with hell and decides to vacation in Los Angeles, was highly offensive to and protested by the Christian right. The American Family Association and One Million Moms both started online petitions to cancel *Lucifer* strictly on moral grounds. The show, both groups argued, "will glorify Satan as a caring, likeable person in human flesh" and fundamentally "mischaracterize Satan." The show, and the network that sponsors it, will "[disrespect] Christianity and mock the Bible." Most of their claims were moral judgements, not based on empirical evidence.[2] Perhaps the one issue that could be empirically established is whether the character of Lucifer was scripted to be appealing.

Some moral questions need considerable supportive evidence, such as claims suggesting that using Twitter at all is bad because it increases the possibility of death. Here, the statement of fact—"increases the possibility of death"—needs to be supported by evidence. However, after doing several literature reviews of different databases, the only evidence found for increased fatal-

1. And the research that has been done, of course.
2. Lucifer is bad (by definition, evil). Portrayal of Lucifer as appealing will decrease the perception that Lucifer is bad. People need to understand that Lucifer is evil to resist his temptation. Therefore, portraying Lucifer as charming is bad. There is no doubt that the television show intended Lucifer Morningstar to be appealing. The Fox press release described Lucifer as "Charming, charismatic and devilishly handsome." Katherine Sangiorgio, "DC's *Lucifer* Pilot Leaks Online Today," Legion of Leia, August 10, 2015, https://web.archive.org/web/20150814091654/http://legionofleia.com:80/2015/08/dcs-lucifer-pilot-leaks-online-today/.

ities was using smartphones in places that are dangerous—such as taking selfies on a ledge, or using Twitter while driving. In the former case, taking selfies is not using Twitter, but taking selfies in hazardous places or situations where the risk for severe injury is high. In the latter case, the danger, however, was not specifically Twitter, but the more general category of reading and texting while driving, which itself is a subcategory of danger associated with "driving while distracted." Since the evidence doesn't support a general claim that "Twitter use per se increases chances of dying," the moral statement, as written, does not hold and is an invalid argument. However, if the moral question was limited to whether texting when driving is bad, then the moral question is supported, given that the audience accepts the overarching moral claim that death is bad.

Policy analysts extensively depend on research to understand a current situation, and to argue for what solution should be adopted.

A literature review for developing a policy to address the impacts of violence in media would at the very least need to look at:

> How much violence is shown in media (a research question)?
>
> Is the impact severe enough that the violence should not be tolerated in society (a moral question that may need empirical data)?
>
> What should be done (a policy recommendation that should minimize the harmful impacts and maximize the positive impacts)?

Policy developers usually need to use available research, which means that the policy analyst would need to extend findings from a specific research population to the population that would be affected by a proposed policy—which can introduce significant distortions if the policy analyst is not careful. A media study that looked at the top grossing films for aggressive acts, for example, would likely underplay the level and intensity of violence in horror films, since horror films—known to be more likely to show intense and graphic violence than many other genres of film—are far less likely to be in the list of top ten grossing films in any one year. Therefore, we can assume that horror film audiences are systematically more exposed to violence than a study of top-grossing films would indicate. A policy maker interested in the question, "Should the U.S. regulate the level of violence in horror films?" would or should know that study of a list that is more likely to have action and adventure films than horror films would likely underestimate the level and explicitness of violence to which horror fans are exposed.

Further, as with all studies, readers need to identify errors the researcher might have introduced in the method, and how those errors bias the study's findings. Coding categories might be incomplete (content analysis), questions could be biased (survey research), or the instrument measuring subject change might not adequately detect change (experimental method). Some of these are

fatal flaws, meaning that the entire study should not be used; some mean that certain questions or certain coding categories are flawed, but that the rest of the findings are still useful.

Constructing a Literature Review

The cost of getting verifiable findings is increased precision in the research question, which usually means narrowing a general question. In most cases, however, describing "what is" means developing an answer to a general question. To do this, analysts put many studies together, using findings from each individual research project as one bit of an overall picture.

In the following chapter, we will walk through building a short literature review to address the question, "What do we know about violence in media?"

21. An Example: Building a Summary

In this chapter, we will go through one technique—a reasonably useful one—for building a careful critical summary of several sets of research findings on a single question: "How much violence is in popular media?" The question of violence in media is both important and trivial. Trivial, because the summary of research for any policy question—or any discussion of any research question, for that matter—would go through a similar process of incorporating results into a description of what is known. Important, because the question determines which research articles are relevant to include and which are not.

To mimic the actual process, for each article, we will start with the notes taken from the methods and the findings of one of the research papers. Generally speaking, the introduction, the review of literature, and the discussion are not included in the summary.

(Note: The introduction generally establishes why a particular research paper is a) a contribution to an important practical problem, b) an important theoretical or methodological problem, or c) an intrinsically interesting problem. Since the starting point of the summary is "How much violence do popular media show?" the reason why each study selected is relevant to the summary is already implicitly given. The review of literature is essentially hearsay evidence,[1] or what the author(s) of the article feels is important in setting up why they are doing the study. While I frequency use the literature review to look for additional papers to read, I generally go to the referred article itself before using any finding cited in a literature review. The discussion section is essentially where the researcher is reviewing the findings and/or extending the findings to other situations (either theoretical or practical). Since the nonrepetitive portion of the discussion section is essentially hypothetical, information from the discussion section is not valid as findings. Hence, the summary should be careful to summarize information about how the findings were developed (the methodology section) and what the findings are (the results or findings section).

Box 21.1 is detailed notes from "Trends of Sexual and Violent Content by Gender in Top-Grossing U.S. Films, 1950-2006," a content analysis study of top grossing films (Bleakley, Jamieson, & Romer, 2012). The notes start with a check on the soundness of the study, move on to a discussion of study weaknesses, and then to the findings.

1. That is, information in the literature review is the author's summary of what other researchers found, not what the researchers said in their own words. It is always best to directly go to the original research paper.

Box 21.1: Study 1 Overview

Study 1. *Trends of Sexual and Violent Content by Gender in Top-Grossing U.S. Films, 1950-2006*

Publication

Amy Bleakley, Patrick E. Jamieson, and Daniel Romer, "Trends of Sexual and Violent Content by Gender in Top-Grossing U.S. Films, 1950-2006," *Journal of Adolescent Health* 51, no. 1 (July 2012): 73-79, https://doi.org/10.1016/j.jadohealth.2012.02.006.

Method

Population. U.S. films from 1950 to 2016

Sample. The top 30 movies each year from *Variety*'s annual list of top-selling 200 U.S. films (1950 to 2006). Researchers selected every other film in each year's list (total 855 films).

Coders. Twenty-four trained undergraduates.

Code for violence. Violence was defined as "intentional acts (e.g., to cause harm, to coerce, or for fun) where the aggressor makes or attempts to make some physical contact that has potential to inflict injury or harm."

Intercoder reliability. Krippendorff's alpha .73 – .77

Weaknesses

Population. Does not cover changes in film content since 2006. Limited to movies shown to U.S. audiences. Biased toward top grossing films.

Coding categories. Emotional violence was not examined.

Findings

Violent content was high, present in 89 percent of all films.

Films that showed explicit sex and violence were significantly correlated.

Both genders started and received violence. Males started and received violence slightly more often than women.

The proportion of violence started and received increased for both genders over time.

Discussion of the Study

Bleakley et al. selected samples from *Variety*'s annual list of top-selling movies. *Variety* uses data compiled by Nielsen EDI and Rentrak Theatrical figures, two of the most reliable sources of data for television viewing and box office (movie) revenue. *Variety* is *the* premier source of entertainment industry news. Industry leaders closely watch it, which suggests that errors in the lists would be noticed and corrected.

Further, since the researchers sampled the most popular movies, they also captured the movies with the widest audience attention, which, arguably, have the most general impact. (Films that appeal to smaller niche audiences are—by definition—less likely to be sampled.)

Box 21.2: Summary of Findings, First Draft

A study of top-grossing movies from 1950-2006 found that 89 percent had violent content. Male characters were slightly more likely than women to deliver and to receive violent acts. Violent acts increased each decade after 1960 for both genders. Films that showed explicit sex were also more likely to show violence. (Beakley, Jamieson and Romer 2012.)

The first look at the research findings should be an answer to the main question of the research summary—in this case, is there violence in media? Looking specifically at Bleakley et al., the reader can say—with a fair degree of confidence—that yes, there is violence in top-grossing films and, in fact, most of the films studied were violent (89 percent). Including the words "top-grossing films" in the write-up also gives a fair clue to the careful reader on *how* the sampling limits findings (see Box 21.2).

The next step is to look for interesting details that show the scope of the violence, including the differences in other relevant factors, such as gender and time. (Both genders started and received violent acts, males slightly more often than women, and all violence increased over time.)

This particular study included other findings which were less relevant to the major question. Generally, only findings relevant to the major question of the summary are included.

Box 21.3: Study 2 Overview

Study 2. *Violent Frames: Analyzing Internet Movie Database Reviewers' Text Description of Media Violence and Gender Differences from 39 Years of U.S. Action, Thriller, Crime, and Adventure Movies*

Publication

Jordy F. Gosselt, Joris J. Van Hoof, Bastiaan S. Gent, Jean-Paul Fox, "Violent Frames: Analyzing Internet Movie Database Reviewers' Text Descriptions of Media Violence and Gender Differences from 39 Years of U.S. Action, Thriller, Crime, and Adventure Movies," *International Journal of Communication* 9 (2015): 547-567, https://ijoc.org/index.php/ijoc/article/view/2921.

Method

Population. U.S. crime films

Sample. Synopses of crime movies published in IMDB from 1973 to 2011 using the genres action, thriller, crime and/or adventure. Movies with "unrealistic violence" were excluded—science fiction, horror, animation, sports (the violence is not the goal), westerns (the violence is historical), and war (the violence is politically endorsed).

Of the 8,932 movies within the realistic violence category, only 16 percent had more than three lines of text describing the movie, leaving a sample size of 1,396 usable synopses.

Coders. Two coders on 10 percent of the entire sample

Code for Violence. The codes for the types and severity of violence were based on data from the FBI Uniform Crime Reports (2009). The categories were as follows: Gun assault, blade assault, physical assault, projectile assault, rope assault, vehicular assault, chemical assault, environmental assault, explosive assault, forced drug use, sexual assault, and unknown assault. In addition to the type of violence, the researchers looked at severity—"light" (e.g., single hits and slaps), "severe" (bloody injury), "lethal," or "unknown."

Intercoder reliability. Cohen's Kappa coefficient = .805

Weaknesses

Sample. One weakness is that the synopses are one step removed from the movie itself and may not accurately reflect the level of violence in the film. Further, writing synopses and assigning genres in developing IMDB content is a volunteer process similar to contributing to Wikipedia.[2] The quality and accuracy of the descriptions can vary widely from reviewer to reviewer, and the researcher does not know whose synopsis is accurate and whose is not. Also, not all movies were reviewed: the number of movies with synopses ranged from a low of 10 percent to a high of 40 percent of all movies released a given year. Again, the bias introduced is not known.

The researchers were primarily concerned with the level of realistic violence, i.e. the violence "that a viewer might actually encounter or read about in the news." They classified action, thriller, crime, and adventure as realistic violence, and excluded genres with "unrealistic" violence, including science fiction, fantasy, horror, animation, sports, westerns, and war movies. The analysis then did not reflect the level of types of violence in a substantial subset of all movies.

Coding categories: Emotional violence was not examined.

2. "Plots," IMDb Help Center, archived December 22, 2018, https://web.archive.org/web/20181222181356/https://help.imdb.com/article/contribution/titles/plots/G56STCKTK7ESG7CP?ref_=helpms_helpart_inline#.

Findings

Violence found

Firearms (39.5% | n = 1,335),

Physical violence (24.8% | n = 839),

Bladed weapons (8.4% | n = 283).

Other, including sexual assaults or violence with chemicals, vehicles, illicit drugs, ropes, and explosives, ranged from 4 to 124 counts (13.7% | n = 464).

Harm

Lethal harm. 54.9% (1,855 acts); Nonlethal: 41.1% (1,387 acts)

Lethal violence type:

Gun violence (67% | n = 898) were fatal shootings)

Blade assaults (58% were lethal| 165 out of 283)

Environmental violence (73% | 91 out of 124), chemical assaults (51% | 35 out of 69), explosive assaults (93% | 63 out of 68), vehicular violence (53% | 31 out of 59), projectile assaults (56% | 29 out of 52), and rope assaults (100% | all 39)

Physical violence (10% lethal l 10% (84 out of 839)

Gender and Violence

Male victims (79.2% | n = 2,676) and perpetrators (80.0% | n = 2,704)

Female victims and perpetrators (13.1% l n = 443 for both)

Male to male violence (63.1% | n = 2,132)

Female perpetrators to male victims (10.3% | n = 348)

Female to female violence (2.2% l n = 74)

Male victim probability of lethal violence decreases significantly over the study period (Wald = 4.838, $p < .001$)

Other

Gun use does not vary from 1973 to 2011.

Introducing New Material into a Research Summary

Each additional research study should introduce new findings, which will either reinforce, expand, or limit the scope of what can be said about a particular topic. The next study's research question, "What is the level of violence in crime films set in the current world (i.e., excluding historical, science fiction, war, sports, and horror movies)?" is both clearly relevant to the policy question,

but from a more limited population. Therefore, we automatically know that the second study will likely underestimate violence in media by not including several genres of film known for having violent content (e.g., horror films, war films) as well as missing violence in other types of films (e.g., romance films) that are more likely to include emotional violence. (Note: the study method in Box 22.3 only tracks physical violence; emotional violence is not included.) Since errors run in one direction—that is, underestimating violence—the careful reader can say that the overall violence is probably more prevalent than what the study suggests.

The Gosselt et al. study (2015) sampled a different population (coding synopses of crime films) than the Bleakley et al. study (top grossing films) (2012), but similarly to Bleakley et al., the synopses of films also likely underestimated the actual acts of violence, given that synopses writers are more likely to overlook a violent act that is not central to the plotline than to introduce violence that did not appear in the film. Again, a literature summary should include enough information that the careful reader can determine bias introduced in the sampling (See Box 21.4). This information is contained within the initial description of the study sample "an analysis of synopses of present-time crime genre movies." The researchers obviously will not tell the reader that the population they studied is not the population that the reader is interested in; readers will have to do that work on their own.

In terms of what findings to include, the literature review should include a description of the level of violence found (the main answer to the literature review question) and additional specific information relevant to the main question. Any of the following would qualify as fleshing out the main question:

> Gun violence was, by far, the most common type of violence (54 percent) and guns the most lethal (67 percent were fatal shootings).
>
> Both men and women were aggressors and victims, but men were far more likely to be the victim (79 percent) and the aggressor (80 percent). Women were aggressors or victims 13.1 percent of the time for each category.
>
> More recent films were less likely (slightly) to have nonlethal violence.

A rewritten literature review (see Box 21.4) combining both studies should first state the new basic answer to the guiding question, and then point out interesting supporting details from both studies. (The words retained from the description of the study in Box 21.2 are in black, and additional information is in green.)

Box 21.4: Summary of Findings, Second Draft

How to Read: Original text is in black. Additions from the *2nd study are in blue and italicized*. [Transcriber's note: source text is referenced in brackets after text]

Major films consistently show a relatively high level of violence both in scope (89% of all top grossing action films from 1950 to 2011 (Bleakley et al 2012)[study 1] *and in number of violent acts overall (Gosselt et al., 2015). [study 2]* Men were more likely to deliver and receive violence, and in one study far more likely. Eighty percent of all violence in crime films were male on male (Bleakley et al 2012). The findings were mixed on whether violence increased over time with one study reporting increasing violence for both genders (Bleakley et al., 2012) [study 1] and *another indicating that male deaths from gun violence decreased significantly over time (Gosselt 2015).* [study 2]

The findings reported should support or qualify the major claim of the topic sentence and not wander off onto other topics. The exact findings used for supporting details will probably vary from writer to writer. For example, Box 21.3 included evidence that supports the claim that movies show a consistently high level of violence, but did not discuss guns or blades versus physical contact (hands, feet etc.). That information would have been fine to include because it helped characterize the violence, but not absolutely necessary. Including that information, however, would be absolutely necessary if the main question for the policy analyst was, "What is the level gun violence in media?"

Looking broadly at the main question of violence in media, the two studies discussed so far only cover one medium (movies) and one particular type of violence (physical). As such, the studies' findings are an incomplete answer. The two additional studies broaden the scope to include other types of media (music videos and TV reality shows) and other types of violence (emotional/psychological violence).

Box 21.5: Overview of Study 3

Study 3. *Violence in Music Videos: Examining the Prevalence and Context of Physical Aggression*

Publication

Stacy L. Smith and Aaron R. Boyson, "Violence in Music Videos: Examining the Prevalence and Context of Physical Aggression," *Journal of Communication* 52, no. 1 (January 2002): 61-83, https://doi.org/10.1111/j.1460-2466.2002.tb02533.x.

Method

Sample. A composite week of music video programming across 20 weeks. (Data gathered during the 1996-1997 television season.) The researchers randomly sampled three popular (at the time of the study) music video channels: Black Entertainment Television, Music Video Television, and Video Hits-One.

Coders. Fifty-six trained undergraduates

Code for violence: Same coding scheme (as above) for violent acts

Intercoder reliability. The intercoder reliability ranged from .67 to 1.0. The type of intercoder reliability was not included.

Coding variables below .7 are:

Pattern of punishment for bad characters = .67

Pattern of punishment for good characters = .68

Additional coding approaching low reliability:

Pain = .70

Depicted harm = .76

Likely harm = .77

Weaknesses

A significant problem with the sample (in terms of applying the findings to a current situation) is that the study was conducted using data from 1997 (over twenty years old) and most of the shows sampled have substantially changed format since the sample was drawn. MTV has drastically changed its programming, BET has transferred its music programming to branded sister networks, and VH1 primarily shows reality television shows. Furthermore, music videos are now most commonly distributed over the web (e.g., YouTube), which has different content restrictions than television.

Coding categories. Emotional violence was not examined.

Intercoder reliability. The intercoder reliability is too low for two of the variables. These are "pattern of punishment for bad characters" (.67) and "pattern of punishment for good characters" (.68).

Findings

Across three channels, 15 percent of all videos in a given week had one or more acts of physical aggression. Of these, 80 percent had one violent interaction, 17 percent had two, and 3 percent had three or more violent acts.

The aggressor was most likely to be an adult (96 percent), male (78 percent), and Black (56 percent). The victims were also most likely to be adult, male, and Black. Males were more likely to be targets of violence in rap (84 percent) or rock (89 percent). The violence differed significantly by music type, with rap videos at 29 percent, rock at 12 percent, R&B at 9 percent, adult contemporary at 7 percent, and other at 9 percent. Just under one-third (32 percent) of all the violent interactions in music videos involved lethal violence that would result in serious physical harm in the real world.

Introduction of a Third Study into a Summary of Research Literature on Violence in Media

Box 21.6: Summary of Findings, Third Draft

How to Read: Original text is in black, additions from *the 2nd study are in blue and italicized*, and additions from the 3rd study are in red and underlined. [Transcriber's note: source text is referenced in brackets after text]

Research [study 3] consistently found violence in media, although the amount varied widely from a low of 15% in music videos (Smith and Boyson 1997) to a high of [study 3] 89% of all realistic crime films (Gosselt 2015). [study 1] *Major films consistently show a relatively high level of violence both in scope (89% of all top grossing action films from 1950 to 2011, Gosselt et al., 2015) and in number of violent acts overall (synopses of present-time crime genre movies from 1993-2011 described an average of 2.4 violent acts per film, Bleakley et al., 2012).* [study 2] All studies showed that men were more likely to deliver and receive violence [study 1], *and in one study far more likely (Bleakley et al., 2012). Eighty percent of all violence in crime films was male on male violence.* [study 2]

The studies of films[study 1 and 2] *were mixed on whether violence is increasing over time, with one study reporting increasing violence for both genders (Bleakley et al., 2012), and another indicating that male deaths from gun violence were decreasing significantly over time (Gosselt 2015).* [study 2] A study of music videos sampled in 1997 from three channels that showed popular music videos (Smith and Boyson 2002) also showed that both victims and perpetrators were disproportionately likely to be adult and male. The videos had fewer acts of violence and were less likely to show fatal events. Just under a third (32%) of the videos had lethal violence, defined as violence that would result in serious physical harm in the real world. [study 3]

The difficulty of adding the findings for the study on media violence, is how to handle questionable findings. For the first time, the study is both old and problematic. First, the authors do not discuss how television channels select the music videos they show, so the direction of any sampling bias on the part of the TV producers is unclear. Second, the data was collected in 1997. There is no substantive reason to believe that music videos are the same now as they were then. Each of the television channels listed in the methods section of the research paper, "Violence in Music Videos," have substantively changed their programming, and most music videos are on distribution channels (e.g., YouTube) that have far weaker content restrictions than mainstream television. So, what to do? Should you throw the data out because it is flawed? Well, no. It is not flawed enough that the research produced has no value. The study data shows that in 1997, music video violence on three channels (that were the major music video channels at the time the research was done) was relatively low. This is a valid finding for over twenty years ago—even though it does not demonstrate the incidence of violence in current music videos. A *good literature review must include all sound relevant research studies, whether the studies agree with the majority of the rest of the findings or not.*

Turning to the specific findings, the music video findings significantly weaken the claims that can be made about violence in media. Instead of saying that there is a uniformly "high level of violence," the overall assessment will need to be softened to accurately reflect the additional knowledge (see Box 21.6, in red) that one study showed some (but not much) violence. The writer also included the year of the study, so that a careful reader will know how dated the research is.

Introduction of a Fourth Study into a Summary of Research Literature on Violence in Media

Box 21.7: Study 4 Overview

Study 4. *Surviving Survivor: A Content Analysis of Antisocial Behavior and its Context in a Popular Reality Television Show*

Publication

Christopher Wilson, Tom Robinson and Mark Callister, "Surviving *Survivor*: A Content Analysis of Antisocial Behavior and Its Context in a Popular Reality Television Show," *Mass Communication and Society* 15, no. (February 2012): 261-283, https://doi.org/10.1080/15205436.2011.567346.

Method

Sample. Seven seasons of *Survivor* (92 episodes). Reunion shows and show teasers were excluded.

Coders. Two trained researchers

Code for Violence

Antisocial acts consisted of four mutually exclusive categories:

Theft: Taking another person's property without that person's consent or knowledge.

Verbal aggression: Any nonphysical act that places another person under duress with the intention to pressure, constrain, or persuade in a noxious manner, or any hostile remark meant to diminish another's self-image or cause psychological harm.

Minor aggression: Physical aggression resulting with minimal harm or no harm (e.g., slapping, punching, and kicking).

Deceit: Misleading for personal gain (e.g., fraud, cheating, and lying).

Intercoder reliability. Krippendorff's alpha = .85.

Weakness

Limited sample: One specific type of television show (reality), and one particular show within that category (*Survivor*).

Findings

There were 4,207 antisocial acts at a rate of 45.7 acts per hour. Indirect aggression and verbal aggression were the most frequently occurring types of antisocial behavior.

The Wilson, Robinson, and Callister (2012) study on the television reality show, *Survivor*, extended the definition of violence to include emotional aggression. This study examined multiple episodes of a single television reality show and found multiple instances of verbal, deceit, and indirect aggression (defined as acts that take "place behind the victim's back") (Wilson, Robinson, and Callister 2012).

The addition of this study means that the topic sentence would need to include both physical and emotional violence, and specific findings about the prevalence and kind of antisocial violence would need to be added in the evidence section (see Box 21.8; the additional changes to the literature review are in brown).

Box 21.8: Complete Literature Review

How to Read: Original text is in black, additions from *the 2nd study are in blue and italicized*, additions from the 3rd study are in red and underlined, additions from the **4th study are bolded and highlighted in yellow**, and *summary conclusions are italicized and underlined and highlighted in gray*. [Transcriber's note: source text is referenced in brackets after text]

Research consistently found violence in all media studied [study 4]– music (15% of all music videos, Smith and Boyson 1997), [study 3] film (89%, *Gosselt 2015* [study 2] and Bleakley et al., 2012), [study 1] and **all reality shows (100%, Wilson, Robinson, and Callister 2012).** [study 4] With the exception of a 20-year-old music video study, [study 3] **all other studies showed high amounts of violence (between 54% and 100%) in television and film.** [study 4] The most violence reported was from a 2015 study of all top grossing action films from 1950 to 2011, (Bleakley et al's 2012 study, [study 1] a finding supported by *Gosselt et al.'s study of synopses of present-time crime genre movies from 1993-2011 which described an average of 2.4 violent acts per film (Gosselt et al., 2015). [study 2]*

Most studies showed that violence was increasing over time for both men and women. [study 4] The outlier for overall violence, music videos, still showed violence in just under a third of the videos. [study 3] **Studies that covered multiple years all showed increases in violence over time (Bleakley et al 2012; Gosselt 2011, Wilson et al 2012),** [study 4] *although, one reported significant reduction in one category – male death from gun-related violence (Gosselt 2015).* [study 2]

Men were involved in the vast majority of physical and a slight majority of emotional violence. [study 4] Eighty percent of all crime film violence was male on male (Bleakley et al 2012). [study 1] Both victims and perpetrators in music videos were disproportionately likely to be adult and male (Smith and Boyson 1997). [study 3] **However, men were only slightly more likely than women to indulge in emotional violence (males = 54% of 3605 incidents, Wilson et al. 2012).** [study 4]

The overall findings suggest a persistent pattern of violence (all studies), with high levels of lethal violence (film and music videos) increasing over time (film and reality shows). The studies, however, do not cover digital distribution modalities – YouTube, Instagram, Snapchat – channels that have distinct distribution patterns and generally fewer restrictions suggesting that the actual level of mediated violence could be more extreme. [summary conclusions]

Finally, the writer should end the summary with any overall assessment that the writer feels is necessary to highlight for the reader, including important limitations of the research findings. For

example, the research studies examined do not include all media. In particular, they do not cover digital media—YouTube, Instagram, or Snapchat. This means that the studies have ignored important distribution channels that systematically have fewer restrictions than traditional media. Since it is probable that the content of less regulated media is more violent, the lack of digital media studies is an important limitation that would bias the overall results. Overall, the literature review should give the reader an accurate sense of the answer to the major question, "Research consistently found violence in all media studied," and enough evidence that the reader would agree with the topic sentence based on the evidence alone (Box 21.8) without actually reading the topic sentence. There is no specific, hard-and-fast rule as to which exact topic sentence the writer can choose to develop and what evidence the writer chooses to include, but there are some shared guidelines for what is acceptable in the topic sentence and what is not. No one, that is, *no one*, who is a rational reader can finish all four studies and say that the media showed no violent content. A writer who suggests no violent content would be delusional or lying. A writer who suggests that movies only showed limited violence would be misrepresenting the level and seriousness of lethal violence in the movies. Anyone in this culture who does not accept that lethal violence is severe is fundamentally detached from shared social reality.

Writers, however, have some choice on secondary research themes, such as:

> The interaction between violence and gender (covered in most studies),
>
> Whether or not the violence shown has consequences (an important factor in how likely viewers will model behavior they see in media and covered in some studies), and
>
> The interactions between violence and race (an important issue in social justice and covered in most studies).

Any or all of these themes would be acceptable subcomponents of the larger theme of violence in media.

The research-by-research rewrite of a research summary is not necessarily the only method to use to construct a summary, given that people have different writing strategies. However, all summaries should have the following structure and inclusions:

a. For each sound research paper, the findings of the paper relevant to the main question should be presented.
b. The literature summary should start with a topic sentence that answers your main question: What did the research papers find?
c. The rest of the paragraph should include pertinent details from individual research papers. The summary should include enough pertinent details that the reader would be able to develop a topic sentence strikingly similar to the actual topic sentence just from

the details alone (without reading the topic sentence).

d. The summary should include shadings and findings that qualify and/or contradict the main summary sentence (topic sentence) of a summary paragraph or paragraphs.
e. The careful reader should be able to pick out where the population in each study differs from the population of interest in the summary. (This information, while important, is the most likely to be missed in extant summaries.)
f. The paragraph(s) should end with a description of where the evidence is incomplete or missing.

22. Conclusions

Social science—for that matter, physical science as well—is above all a way of knowing. And scientific knowing is based on methods.

Each method—whether it has been focused on learning what people are willing to tell you about themselves (surveys), or examining the artifacts that people produce (content analysis)—has both strengths and weaknesses. The strengths, of course, are what kinds of knowledge the methods can produce. That is, can the method tell you what impact something has on people (experiment), or is the method collecting a range of statements from a specific population (survey)? The weaknesses are the specific ways that the findings are limited. (For example, research on the experience of sitting in a movie theater would be limited if the researcher only talked to short people; the troubles that tall people have would not be addressed because no tall people were interviewed.)

The rules for each method discussed in the main body of this book essentially allow you to tell the generalizability and reliability of each individual research effort. With each different method, the rules differ for determining how greatly the people or objects studied resemble the overall population (generalizability) and how likely it is that another study using the same method on the same population will deliver the same results (reliability).

For example, both the strength and weakness of experiments is that the experimenter controls exactly who the subjects are and what the subjects do. The researchers can rule out other potential explanations, leaving the one thing the experiment manipulated (the treatment) as the only explanation left.

At the same time, the experimenter is also creating a highly artificial situation. Because the experiment is artificial—a laboratory situation in most cases—the experimental results may have low ecological (real-world) validity. In general, the more artificial the experiment is, the greater the likelihood that other factors—not present in the experiment—could change what happens in the world. That does not mean that you can "throw out" the results that you do not like because the experiment is artificial (throwing out research whose findings you do not like is an example of motivated reasoning), but it does suggest that you interpret the results with some degree of caution.

Take, for example, an experiment in which the researchers are trying to check whether using a smartphone while crossing a busy street is risker than not having a smartphone. The researcher considers using two different methods. In one, the researcher sets up a virtual scene of crossing the road in the laboratory. He has his experimental subjects uses their smartphones while deciding

when to cross the street, while the control subjects are simply carrying their smartphones in their hands. The researchers described the experiment setting as follows:

> "Traffic flowed in a bidirectional manner on three computer monitors arranged in a semi-circular manner in front of participants. Participants stood on a wooden "curb," watched the traffic in a first-person point of view, and stepped down off of the curb when they deemed it safe to cross. The perspective then changed to third person point of view, allowing participants to see whether they made it across the street safely. The speed at which a participant crossed matched that of their average walking speed recorded earlier in the session. Ambient and traffic noise was delivered through speakers."
>
> Schwebel et al.[1]

In the second experiment, the researcher could videotape people naturally crossing a street in the course of their daily lives. The researcher could then used his video record to compare how safely smartphones users were in real-world conditions, compared to nonusers.

The first experiment, in the laboratory, has fairly low ecological validity—stepping off of a wooden block while looking at three computer screens isn't the same as crossing a real street. It's possible that people in the laboratory knew (either consciously or at some low, barely noticeable level of awareness) that they were a lot safer in a laboratory than they would be sharing a walking space with two to eight thousand pounds of moving plastic and steel. Perhaps that knowledge relaxed them and make them a little bit less careful. It's possible. But people in the second experiment are sharing the space with trucks, cars, and SUVs. Because the second experiment was recording real people really crossing the road, this experiment has high ecological validity.

But the first experiment still showed that people with smartphones were more distracted and crossed the (albeit virtual) street less safely than people without cellphones. Those results do not go away just because the ecological validity is lower. The research has shown that there is some potential for concern, and that concern needs to be explored and dealt with. Future research will either continue to show a safety risk for using cell phones while crossing roads or not. But at some point, a decision will need to be made that the evidence is "good enough" to warrant some policy decision.

Reading and critically evaluating scientific literature give the reader power to critically evaluate research when that research is important to them—whether the importance is understanding a subject or making a decision. One of those powers is essentially understanding how to interpret

1. David C. Schwebel et al., "Distraction and Pedestrian Safety: How Talking on the Phone, Texting, and Listening to Music Impact Crossing the Street," *Accident Analysis & Prevention* 45, no. 2 (March 2012): 266-271, https://doi.org/10.1016/j.aap.2011.07.011.

the results according to population differences. Suppose you—or someone you care deeply about—was diagnosed with stage four liver cirrhosis from Non-Alcoholic SteatoHepatitis (NASH). The doctors and the web tell you that a survival rate for people with a stage four diagnosis is one to five years at diagnosis. However, the population data from which the doctors are drawing is all patients with the diagnosis of stage four cirrhosis—alcoholics and non-alcoholics together. The treatment potential for the two groups is substantively different. The non-alcoholics are at least potentially eligible for a liver transplant, while active alcoholics are not.

The survival rates of people waiting transplant are also important to critically evaluate. The medical profession uses MELD scores to determine liver function.[2] MELD scores range from six to forty, with forty being essentially a death sentence within a matter of months, if not days. Technically, however, the MELD score predicts the likelihood of death over the next ninety days of illness. A MELD score of twenty-one, for example, gives a patient a one in five chance of dying within the next three months. However, the three-month mortality range is commonly given at MELD score ranges of below ten, ten to nineteen, twenty to twenty-nine, thirty to thirty-nine, and forty. Therefore, logically speaking, an individual with a score of twenty-one has less of a chance of dying within three months that the MELD score estimate would suggest. In addition, people with strong family support who are younger and female are also less likely to die than isolated, older males. In other words, the more detailed your understanding of the population, the more you can assess individual risk.

Essentially the judgment call that policy makers make—always—is on how great a risk is. Would you give up watching your favorite television shows because one study in a laboratory with thirty college students in Holland showed that students who watched the show were less likely to open up a door for a stranger? Probably not—that would be betting the farm on a very small study done in a different culture. Would you continue to argue that smoking is not harmful to your health after fifty years of research, with thousands of studies that have shown smoking increases cancers, heart problems, skin problems, lung problems, and a host of other health-related problems? Only if you were working for cigarette companies, and probably not even then. To ignore the weight of the evidence that we now have about the dangers of smoking is simply willful ignorance. Even if an individual decides that the pleasure (benefit) from the next cigarette is worth the risks, the vast, *vast* majority of smokers and nonsmokers today accept the evidence that both this cigarette and the next pose a health risk.

2. MELD is an abbreviation of Model for End-Stage Liver Disease. The Meld score ranges from six to forty and is based on the results of several laboratory tests that determine creatinine, bilirubin, serum sodium, and clotting factors (INR). Model for End-Stage Liver Disease (MELD) for ages twelve and older. Kiran Bambha and Patrick S. Kamath, "Model for End-stage Liver Disease (MELD)," in *UpToDate*, edited by Bruce A. Runyon, accessed July 2023, https://www.uptodate.com/contents/model-for-end-stage-liver-disease-meld

At some point, policymakers—and we are all policymakers—need to weigh what policies are appropriate given what we know now about how our current practices—whether watching television or surfing the net during lectures—affect us and the people around us, a process that includes—but is not limited to—judging the soundness of the available research and how much risk an individual or a society is willing to tolerate.