Research Methods in the Social Sciences



Dr. Marina Klimenko

SENTIA PUBLISHING

Research Methods in the Social Sciences

Dr. Marina Klimenko

Copyright © 2023 by Dr. Marina Klimenko

Sentia Publishing Company has the exclusive rights to reproduce this work, to prepare derivative works from this work, to publicly distribute this work, to publicly perform this work, and to publicly display this work.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the copyright owner.

Printed in the United States of America
ISBN 978-1-7359340-2-0

Table of Contents

Chapter 1: How It All Began 1
Chapter 2: Principles of Modern Science 15
Chapter 3: Thinking Like a Scientist 33
Chapter 4: Content Analysis 65
Chapter 5: Observational Method 95
Chapter 6: Experimental Research 109
Chapter 7: Quasi-Experimental Design 133
Chapter 8: Understanding the Logic of Statistics and Describing Data Like a Scientist 147
Chapter 9: Testing a Statistical Hypothesis with SPSS Like a Scientist 173
Chapter 10: Writing Like a Scientist 229
Chapter 11: Becoming a Critical Thinker and a Smart Consumer of Science 255

Chapter 1: How It All Began

Introduction

"The important thing in science is not to obtain new facts, but to discover new ways of thinking about them."
— *Physicist William Lawrence Bragg*

Psychology is the study of human behavior and mental processes. It is also a scientific discipline as it uses *science* to test theories and make predictions. The main goal of this textbook is to introduce you to the most common *scientific* methods used in the psychological field and to show that psychology is a scientific discipline in its own right. Throughout the chapters, I will use the words *science* and *research* interchangeably, given that psychological research is based on scientific principles, and thus, the two words share meaning. However, defining science may not be a simple task. To fully grasp and appreciate it, you will need to go through the entire scientific process. For now, let's simply define science as a process of inquiry that involves both systematic observations and critical thinking. We can further break down this process into the following stages: Asking research questions and formulation of hypotheses, data collection and analysis, interpretation and dissemination of results and, finally, peer-review and refinement of knowledge. Walking through every stage of this process will give you a fuller appreciation for both the field of psychology and scientific thinking. A good scientist is the one who "believes and acts entirely on the basis of evidence and reasons" (Siegel, 1989). Although, this is not always the case (scientists are also humans), scientists should leave their self-interests and opinions at the door, and strive to be "dispassionate seeker of the truth" (Siegel, 1989).

The goal of this chapter is twofold. First, I will demonstrate that science, as a method of gaining knowledge, did not appear out of nowhere or as an accident, but, in my judgment, has been with us, as part of our human intellect, from the beginning of our human existence, even if it was then in its most primitive form. Since the conversation about human nature is beyond the scope of this textbook, I will highlight only a couple of important human traits that have played a decisive role in the development of knowledge throughout our history. The second goal is to illustrate that the accumulation of knowledge did not begin with the scientific revolution; there are other methods that humans have employed in the past and continue to use today, for example, observation, logical reasoning, and learning from more-knowledgeable others. Still, no method of acquiring knowledge can compete with the scientific method in terms of reliability. Finally, by outlining the development of scientific thinking, I hope to help you gain a deeper understanding of the philosophical underpinnings of modern science.

1.1 Accumulation of Knowledge

Although the beginning of modern science "officially" dates back to the period of the Renaissance, which roughly lasted between the 14th and the 17th centuries, the scientific method, as we know it, wasn't invented suddenly, but rather, gradually became a method of understanding the world over millions of years. The intellectual precursors of modern science have been part of our existence from the very beginning, starting with pre-historic man, and, some can even argue, thus make us a unique species. While researchers don't have a consensus on what those special traits or behaviors are, making this a debatable topic, several behavioral patterns and skills have been studied enough to support the belief that they played an essential role in our cultural and intellectual evolution.

Humans are remarkably good at solving problems to make life easier. This is quite obvious if we simply take a look at all the technological advances that we have made to improve our living. Problem solving skills are also important in scientific endeavors. However, problem solving is not a uniquely human trait. For example, our closest ancestors, chimpanzees, are also fairly good problem solvers.



Chimpanzees employ problem-solving skills daily in their natural habitat. For example, they form alliances, cooperate during hunting, use stone hammers to get kernels out of their shells, and, of course, perform many other tasks collaboratively or alone, depending on the nature of the problem (e.g., Sakura & Matsuzawa, 2010; Watts & Mitani, 2002). Yet, given what we have accomplished as a species, our intellectual superiority is unquestionable. One of the striking differences between

us and our closest living relatives is that, unlike them, we don't just stop after the problem has been solved; rather, we try to preserve the solution and share it with the others—we pass the information we acquire onto others orally, in writing, or by modeling the behavior, a process known as *cultural transmission*. We preserve information, methods and technologies that represent improvement in efficiency (e.g., Dean, Vale, Laland, Flynn, Kendal, 2013; Flynn, 2008). In other words, it isn't just the facts that we are interested in remembering but the process of obtaining them—the 'what' and the 'how.'

There is at least one good reason for why we culturally transmit. It is very efficient—information once learned can be reapplied to address the same challenge without wasting time and energy looking for the answer once again. Furthermore, even if only one person has solved the problem, the information will spread to others—we don't have to individually participate in the problem solving to possess that knowledge. Just think about all the discoveries that we have inherited in some form or shape from our ancestors and didn't have to rediscover. For example, we take for granted the fact that Earth revolves around the Sun, thanks to Nicolaus Copernicus and his book, *The Revolutions of the Heavenly Bodies*, which was published in 1543. In his book, Copernicus proposed a new system of planetary motion, where the Sun, not Earth, was the center of the solar system. However, if the historian David Wootton was correct, it took about 68 years for this fact to be accepted by the public—the idea was too radical for that time and it was only after Tycho Brahe's observation of the nova in 1572, the observation of the comet of 1577, and Galileo's invention of the telescope in 1609, which proved Galileo's discovery of the four moons of Jupiter in 1611, that the geocentric solar system could no longer make sense and was replaced by the Copernicus' system (Wootton, 2015).

Cultural transmission is very adaptive. Some researchers argue that evidence exists for social and cultural transmission of knowledge in non-humans as well. For example, according to researcher Boesch (2003), the use of hammer stones by chimpanzees to crack nuts is actually an improvement over hitting nuts on the substratum by their ancestors. This and other examples, according to Boesch (2003), represent cultural transmission in chimpanzees. Still, even if we grant some cultural transmission to non-humans, it doesn't match the level of the social learning and cultural transmission that happens in humankind. As archaeologist Steven Mithen (1999) stated, "Each generation of chimpanzees appears to begin from scratch on the same old problems as their parents, grandparents and all great-grandparents before them had to learn to solve" (p. 5).

One of the interesting byproducts of cultural transmission is the evolution of prestige, or valuing individuals who possess above-average knowledge of some sort (i.e., models), and either trying to become that individual (i.e., the model) or, as a copier, trying to obtain proximity to the model so that the knowledge or the skills can be copied or learned, according to the researchers, Henrich and Gil-White (2001). For example, they pointed out that in simpler societies where hunting is the primary method of obtaining food, a skilled hunter will be highly valued and will receive higher status and prestige, while other people will try to form close relationship with this individual to

gain some of his/her expertise (Henrich & Gil-White, 2001). Whereas, in more complex societies, with more complex division of labor and ever growing knowledge necessary to perform any given labor, people of more diverse occupations or those who can produce new knowledge (e.g., scientists, inventors) are highly valued and respected.

Prestige, as a strong evolutionary force, can also explain why people go to great lengths, and in some cases risk their lives, trying to leave their mark in history by inventing or discovering something new, as began to happen after the explosion of science (Wootton, 2015). For example, Galileo was accused of heresy and sentenced to house arrest for his beliefs and his scientific work to prove the Copernican's theory of a heliocentric system. At that time, this was an extremely radical idea because it challenged the very existence of God and shook people's entire view of the world. This, however, didn't stop him from continuing his scientific endeavors or publishing more and arguably even more controversial book in 1638, *Dialogues Concerning Two New Sciences* (Hergenhahn, 2001, p. 94).

Clearly, there are certain traits that have allowed human species to excel at cultural transmission and accumulation of knowledge. A couple of them will be discussed next.

Symbolism

The human drive to share knowledge with conspecifics, to accumulate and to build on previous technological achievements, may be that one special human characteristic that gave us an evolutionary edge. It is also what continues to excite many people today, stimulating scientific inquiries. Another important intellectual quality that goes hand in hand with sharing knowledge is *symbolism*, the ability to abstract physical reality into symbols, have reflective thoughts, and, of course, communicate with language (see Greenspan and Shanker, 2006). This intellectual capacity gave humans not only the means by which they could share and accumulate knowledge but also the tools to better understand themselves and the world. Humans could then label their sensory experiences, explicitly discuss their actions, and reflect on their experiences and thoughts. In other words, symbols, such as pictures or words, could be used to formulate ideas, to share them with others, and to make appropriate adjustments to strategies in order to improve technology.

Our first technology dates as far back as 2.5 million years ago, and is known as Oldowan stone tools (see Figure 1). As primitive as they were, making them involved several stages: choosing the right size, shape, and composition of a rock, and hitting it with another rock at a particular angle and the perfect percussion.



Figure 1. *Oldowan Chopper*
Source: Copyright by Curators of the University of Missouri. All rights reserved. DMCA and other copyright information.

It took our ancestors quite some time to improve on Oldowan production, but around 1.6 million years ago, they began making better and more effective tools, known as Acheulean tools (see figure 2). Their unique shapes as hand axes or picks were more versatile and could be used for more complex tasks, such as digging in soil, butchering, skinning, and other chores.



Figure 2. *Acheulean Handaxe*
Source: Copyright by Curators of the University of Missouri. All rights reserved. DMCA and other copyright information.

Between 100,000 and 70,000 ybp (years before present time), humans were also developing more complex social lives and more specialized behaviors. For example, males would specialize in more dangerous activities (e.g., hunting), and females would stay closer to shelter in order to take care of children (e.g., Rossano, 2010). All of these innovations in technological productions and social behaviors required abstraction in action organization, collaborations, and social transmission of knowledge over hundreds of thousands of years. According to many scholars, this had to be accomplished through different forms of symbolic expressions, probably storytelling and rituals, but also through art in the form of cave paintings, carved wood and stone figurines.

Learn more about symbolic formation in humans and how symbolism is different from our non-human ancestors, chimpanzees: http://humanorigins.si.edu/human-characteristics/language (Or go to the website of Smithsonian National Museum of Natural History, choose What Does It Mean To Be Human? and click on Human Characteristics: Language and Symbols)

Prosociality



Despite the fact that our history is filled with wars and violence, we are, surprisingly, empathetic and prosocial creatures. First, we are incredibly good at understanding other people's feelings and thoughts. We can even experience the same physical sensations that another human being is feeling—for example, when we are watching an emotionally charged movie scene we may start crying when seeing the protagonist trying or we may share the laughter.



Psychologists call these sets of skills theory of mind; various aspects of theory of mind begin to emerge gradually as early as in infancy (while some maybe innate) and reach a relatively sophisticated level by the age of 4. Having this incredible emotional and intellectual connection with other humans can explain why we can act altruistically—do things for other people even if this may be against our own self-interests. These traits have also made us effective collaborators—knowing who shares our goals can help us recruit people for their unique talents and ideas and, collectively, we are better equipped at solving any problem. Thus, prosociality may be one of the human traits that set us apart from our closest relatives, chimpanzees: If social learning (i.e., learning from others) is a more efficient and adaptive strategy than individual learning to culturally transmit and improve technologically, our superior social skills most certainly play a key role; being prosocial may also explain our general concern for the well-being of our species and the obligation that many people feel to help the next generations have better lives.

Scientific evidence seems to support the evolutionary importance of prosociality in human intellectual and cultural evolution. For example, researchers, Deen, Kendal, Schapiro, Thierry, and Laland (2012) conducted an experiment. In it, groups of three- and four-year-old children, chimpanzees, and capuchin monkeys were asked to work on a puzzlebox for rewards; the puzzlebox had three stages of solution; the successful completion of each stage depended on the completion of the preceding stage, and each stage had two-solution options. Of course, the tasks were appropriately designed for each group, so that all three could realistically solve them.

The idea behind the three-stage puzzlebox was that it would stimulate some collaboration among the group members, and this would allow the researchers to examine any differences or similarities

among the primates in how they learn and share information in the process of solving the problem. The results of the experiment revealed that prosocial behavior was one of the behaviors that distinguished the human group from the other primates and played the key role in solving the puzzle. For example, only the children exhibited an act of altruism by spontaneously sharing rewards with other children in their group during the solution of the puzzle, an action that kept all group members motivated to keep working on the puzzle. No instances of such altruism were observed among the chimpanzees or capuchin monkeys.

Take-Away Message: Human symbolic expression, prosocial behavior, and accumulation of knowledge are unparalleled in any other species. We have been seeking solutions to our problems since the dawn of time. However, in addition to solving them, we have gone to great lengths to pass our discoveries on to the next generations. Among its myriad impacts, the social and cultural transmission of knowledge has lent itself to building on previous discoveries and making more efficient and effective technologies with every consecutive generation (e.g., Kurzban & Barrett, 2012).

1.2 Transition to Agriculture and the Birth of Great Civilizations

If the evolution of social brains had enabled us to learn from others, living in larger communities with more various people (i.e., various minds), had led to "extracting 'better quality' (more fitness enhancing) information from the minds and behavior of others." (Chudek & Henrich, 2011). You've heard the expression, "two heads are better than one", which means that, collectively, people come up with more and better ideas than they do individually. This commonsense fact can also explain the association that exists between human population density and innovation—where there are more people, more ideas can be generated, and these ideas are easier to spread among other people to be improved and built upon (e.g., Carlino, Chatterjee, & Hunt, 2007). In prehistoric time, there was one event that occurred around 10,000BC which had a profound effect on our human culture, and its population growth, and, in turn, accelerated the acquisition of knowledge and development of more complex technologies—the transition from a hunting and gathering lifestyle to settled living, use of agriculture and domestication of animals, which is known as the Neolithic Revolution. Although researchers are still debating what propelled humans to make this giant leap, there is little doubt that by that time, humans already knew how to manipulate nature to produce food, domesticate animals, and manufacture new tools in order to support their settled communities.

Tribes grew into statehoods and empires, and the growth and transformation of communities had become a catalyst for even more rapid technological innovations. Ancient Egypt is the prime example of the transformation that took place during a relatively short period of time in the Mediterranean region. Ancient Egypt grew into a highly sophisticated and complex society, with

distinctive art, advanced technology, politics, and, of course, religion. Ancient Egyptians cultivated grains, vegetables, and fruits; they were familiar with and used as many as 2000 types of flowers and other aromatic plants and spices in food, during religion practices, and as cosmetics and medicine; they developed various harvesting and post-harvesting tools, and invented hydraulic engineering and irrigation systems; and they were able to control pests, grew gardens, and ventured out into unknown lands (Janick, 2002). Their striking architecture continues to marvel us today with its beauty, massive and complex structures, and perfect designs.



To learn more about the achievements of ancient Egypt, go to The British Museum website: http://www.ancientegypt.co.uk.

Take-Away Message: Transition to agriculture caused a dramatic increase in population and growth of empires, which, in turn, led to a more rapid accumulation of knowledge and technological improvement. Great civilizations, such as the Egyptian dynasties, serve as a testament to the effect of population growth on human intellectual evolution.

1.3 Early Greek Philosophy

Before ancient Greek civilization, human knowledge and technological achievement were already quite impressive. However, human inquiries were limited to practical necessities—for example, domestication of plants and animals was supposed to provide a stable supply of food, and epic tombs and temples were built to make political statements and give space to practice religion. The ancient Greeks began to ask questions and systematically organize their knowledge for one more reason-- because they were curious about the natural order of things (Hergenhahn, 2001, p. 26). Their inquiries ranged from the laws of nature to the nature of knowledge itself.

The etymology of the word *philosophy* is a coupling of two Greek words, *phylos*. meaning "love," and *sophie*, meaning "wisdom." Ancient Greece was a perfect place to embrace the freedom of thought and to have the audacity to question traditional teachings. First, as a trading center for the Mediterranean Sea, Greece maintained constant trade relations with the nearby empires of Egypt, Syria, and Northern Europe, which made social transmission and diffusion of ideas easier and more frequent. In addition, Greeks were not constrained by their religious beliefs because their priests did not have as much political or economic power as in other places. The Greek elite enjoyed the freedom of public debates, deliberations, and arguments. Public debates would take place in every city-state in ancient Greece, in an *agora*, a special public place of congregation (Davernport & Leitch, 2005). The best known was in Athens, where the agora developed into a major center of business and social life by the 6th century BC. Political participation was an important duty of every citizen. The democratic spirit and certain political freedoms spilled over even to people of lower legal and social status. It is important to mention that Greek democracy coexisted with slavery, and slavery was rationally justified by Greek thinkers. Still, freedom of thought, political rights, and the fusion of different ideas played a big role in the birth of philosophy, the precursor to psychology and to science as a whole (Ben, 1908).

Thales (ca. 625–545 BC) is thought to be the first of the pre-Socratic Greek philosophers. Besides pondering the origin and the structure of the universe, many first philosophers, including Thales, were interested in a range of other subjects. For example, Thales studied geometry and by many accounts, discovered that the angles at the base of an isosceles triangle are equal, and that if two straight lines intersect, the opposite angles are equal. Thales was also interested in astronomy, and some claim that he predicted a solar eclipse. The idea of evolution was first suggested by Anaximander (ca. 610–540 B.C.), a student of Thales, and was expanded upon by Empedocles (ca. 495–435 B.C.). Democritus even claimed that humans were made of tiny parts called *atoms* (Hergenhahn, 2001, p. 31).

To learn more about Ancient Greece, go to an Ancient Greece website: www.ancientgreece.co.uk.

Critical Thinking

The ancient Greeks' arguably most important contribution to science was the audacity to openly challenge the status quo—the worldview rooted in religious and supernatural belief. The practice of challenging and questioning "known facts" privately and in open debates became a tradition in Ancient Greece. And the importance of critical thinking in science and innovation is unquestionable in the modern day (Hergenhahn, 2001, p. 47). Although erroneously, early Greek philosophers believed that critical thinking and reasoning were the only methods of discerning knowledge from appearances. Senses, in their view, were deceiving, and thus, could not be trusted if one is seeking true knowledge. Plato referred to his method of questions and answers as *dialectic.* He categorized the world into the sensible and intelligible. What we experience with our senses is the sensible world; it is confusing and doesn't represent true reality. The absolute truth exists in its pure form and can be accessed only through rational thinking. When we apply our intellect, we access the intelligible world, the absolute forms of reality (Dumitriu, 1977, p. 109). Aristotle, Plato's most famous student, had a slightly different view of absolute knowledge. He asserted that knowing involves accessing the sensible world through observations. However, he agreed with Plato that sensory experiences are not true knowledge yet, and that one has to apply reasoning to obtain *universals,* the ultimate knowledge of things.

Critical thinking remains an essential part of a research process and can be defined as "reflective thinking involved in the evaluation of evidence relevant to a claim so that a sound or good conclusion can be drawn from the evidence" (Bensley, 1998). The emphasis is placed on evaluating a claim based on the merits of its evidence. Individuals who embrace critical thinking are generally concerned with seeking truth and thus, are not afraid to question the logic of their own beliefs and to consider alternative ideas and perspectives if evidence dictates. Since at the heart of science is finding truth, critical thinking is necessary when we conduct research; it protects us from our own biases and guides us in every step of the research process, from formulating the right questions and choosing the right methods of data collection and statistical analysis, to refining our working theory based on newfound evidence. Just like a critical thinker, a scientist can be called as someone who is "appropriately moved by reasons" (Siegel, 1989), but in addition to reasons, a scientist must consider the evidence (data).

To better illustrate the cognitive skills involved in the critical thinking process, I will use the model developed by Bloom (1956). This model identifies six hierarchical steps of learning. The basic first step is *knowledge*, i.e., before you can engage in critical evaluation, you have to be familiar with the subject matter. The second step is *comprehension*, understanding the subject at hand. Once you have achieved comprehension, you can apply information as intended. The last three steps are *analysis*, *synthesis*, and *evaluation.* These are the critical thinking levels, which require you to break down the information into its various components (main idea, supporting evidence, etc.), evaluate supporting evidence, and make a conclusion based on the merits of the presented information.

The purpose of this text is to give you the basic knowledge to understand the process of science—the first two levels of learning. My hope is that you can then apply this knowledge to become a critical thinker—analyzing and evaluating information that comes to you before you can accept it. Together, knowledge of science and application of critical thinking skills will make you a better consumer of science, as you will be able to discern bad from good science or personal opinion from evidence-based argument. In the words of the physicist William Lawrence Bragg, your job as both a student and an intellectual human being is "not to obtain new facts, but to discover new ways of thinking about them" —whether you're debating, reading, or writing.

References

Ben, A. W. (1908). *Early Greek philosophy.* London: Archibald Constable & CO Ltd.

Bensley, D.A. (1998). *Critical thinking in psychology: A unified skills approach.* Pacific Grove, CA: Brooks/Cole.

Bloom, B. (1956). A taxonomy of educational objectives. *Handbook 1: Cognitive domain*. New York: McKay.

Boesch, C. (2003). Is culture a golden barrier between human and chimpanzee? *Evolutionary Anthropology 12*, 82–91.

Carlino, G. A., Chatterjee, S., and Hunt, R. M. (2007). Urban density and the rate of invention. *Journal of Urban Economics, 61*, 389–419.

Chudek, M., and Henrich, J. (2011). Culture-gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences, 15*, 218-226.

Davernport, S., & Leitch, S. (2005). Public participation: Agoras, ancient and modern, and a framework for science-society debate. *Science and Public Policy, 32*, 137–153.

Dean, L. G., Vale, G. L., Laland, K. N., Flynn, E., & Kendal, R. L. (2013). Human cumulative culture: A comparative perspective. *Biological Reviews*, 284–301.

Deen, L. G., Kendal, R. L., Schapiro, S. J., Thierry, B., and Laland, K. N. (2012). Identification of the social and cognitive processes underlying human cumulative culture. *Science, 2*, 1114–1118.

Dumitriu, A. (Ed.). (1977). *History of logic* (Vol. 1). Abacus Press: Tunbridge Wells, Kent.

Gabora, L. (2008). Mind. In R. A. Bentley, H. D. G. Maschner, & C. Chippendale, (Eds.) *Handbook of Theories and Methods in Archaeology* (pp. 283–296).Walnut Creek CA: Altamira Press.

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. Evolution and Human Behavior, 22, 165–196.

Hergenhahn, B. R. (2001). *An introduction to the history of psychology*. Wadsworth Thomas Learning.

Janick, J. (2002). Ancient Egyptian agriculture and the origins of horticulture. p. 23–39. In S. Sansavini and J. Janick (Eds.), *Proceedings of the International Symposium on Mediterranean Horticulture Issues and Prospects.* Acta. Hort. 582.55-59.

Kurzban, R., & Barrett, H. C. (2012). Origins of cumulative culture. *Science, 335*, 1056–1057.

Lee, S. J., Srinivasan, S., Trail, T., Lewis, D. & Lopez, S. (2011). Examining the relationship among student perception of support, course satisfaction, and learning outcomes in online learning. *The Internet and Higher Education* (2011), doi: 10.1016/j.iheduc.2011.04.001

Mithen, S. (1999). Problem-solving and the evolution of human culture. *The Institute for Cultural Research Monograph*, 33.

Olsson, O., & Hibbs, D. (2005). Biogeography and long-run economic development. *European Economic Review, 49*, 909–938.

Rossano, M. J. (2010). Making friends, making tools, and making symbols. *Current Anthropology, 51*, 89–98.

Sakura, O., & Matsuzawa, T. (2010). Flexibility of wild chimpanzee nut-cracking behavior using stone hammers and anvils: An experimental analysis. *Ethology, 87*, 237–248.

Siegel, H. (1989). The rationality of science, critical thinking, and science education. *Synthese, 80*, 9-41.

Smith, S., & Young, P. D. (2012). *Cultural Anthropology: Understanding a world in transition* (2nd ed.). BVT Publishing, LLC.

Watts, D. & Mitani, J. (2002). Hunting and meat sharing by chimpanzees at Ngogo, Kibale National Park, Uganda. In C. Boesch, G. Hohmann & L. Marchant(Eds.), *Behavioural Diversity in Chimpanzees and Bonobos*, pp. 244–255. Cambridge: Cambridge University Press.

Weisdorf, J. L. (2005). From foraging to farming: Explaining the Neolithic revolution. *Journal of Economic Surveys, 19*, 561-586.

Wootton, D. (2015). *The Invention of Science: A new history of the scientific revolution.* (EPub). Penguin Random House UK.

Chapter 2: Principles of Modern Science

Introduction

In this chapter we will look at the key players and events that led up to the birth of modern science. We will also discuss the main philosophical principles and stages of the scientific process that set it apart from all other forms of inquiry.

2.1 Alternative Forms of Inquiry

In the previous chapter, we briefly surveyed the history of human intellectual progress. We saw that people made new discoveries, solved everyday problems, and passed down knowledge from generation to generation before the invention of modern science. So why do we treat science as a special form of inquiry? To answer this question, let's first look at some of the common nonscientific methods we casually rely on daily to learn and to understand the world around us.

Unsystematic Observations

People learn by imitating other people and by observing patterns in their environment. Some of the greatest human achievements have come as a result of careful but unsystematic observations. For example, people learned to grow their own food most likely by accident, by first noticing that accidentally disseminated seeds could produce crops under certain conditions. Imitation — observing and copying actions of other people — is still a very important form of social learning, particularly for young children, who begin to imitate people's facial expressions, gestures, and motor actions as early as infancy. Developmental psychologists believe that learning by observing and imitating others an important mechanism by which all humans learn to understand and to share emotions, to master new skills, and to acquire cultural norms. However, this method does have its downsides.

Relying on casual, unsystematic observations is like trying to find a needle in a haystack—you have too much information to sort through in order to get to the important data. Imagine, for example, that you decided to prepare a new dish without using a recipe. Even if you are familiar with the dish's ingredients, you will likely try and fail several times before you successfully prepare the dish. Cooking without a recipe, in this example, is like attempting to discover the solution to your research question without proper guidance, and cooking through simple trial and

error is similar to relying solely on prior individual experiences and unsystematic observations—this method is usually less efficient and more likely to yield erroneous answers.

Another major problem with relying on individual observations is that there is no way to discern whether what we experience is typical and can be generalized to other people and situations, is atypical and unique to one's personal circumstances, or is simply an incident. Despite this problem, we rely on our individual observations to make general statements as if they are facts: for example, we make judgments about people's characters, theorize about causes of behavior, and make predictions. As a humorous illustration, recall the HBO sitcom "Sex and the City." The story portrays four single female friends, each with her own unique perspective on life and relationships. Samantha is a woman who loves sex and doesn't want to be tied down by love or a long-term relationship. Miranda is a cynical lawyer who tries to counter-balance her independence with being in a relationship. Charlotte, the complete opposite of Samantha, believes in love and the *happily ever after.* Finally, the main character, Carrie, is a combination of all three of her girlfriends—she is smart and witty, independent and cynical, yet romantic and hopeful. They mean well when they give each other advice, yet their theories about men are often flawed because they are as different as their personalities and experiences, making them unreliable and biased.

Relying on unsystematic observations can also lead to oversimplified and stereotypic thinking, especially given the fact that people are habitual contingency detectors--they tend to view events that happened to co-occur in close proximity as causally related (i.e., thinking that one event has caused the occurrence of the other) even if their co-occurrence was purely random. From an evolutionary perspective, this tendency is a very effective survival tool; it helps us quickly learn about dangers in our environment and avoid them in the future. For example, eating and getting sick shortly thereafter are seen by most people as two causally related events; and in many cases this may very well be the case, teaching us to stay away from food that caused the illness. However, other times we see causal connections that are not there, and only because the happen to coincide in time.

For example, in the experiment by Hill, Lewicki, Czyzewska, and Schuller (1990), subjects were presented with pictures of different faces, which were purposefully paired with some made-up personality types. The association between the faces and the personalities was very subtle, and the subjects were explicitly instructed *not* to pay attention to the faces and focus only on the personality profiles. In other words, the subjects were not consciously aware of the subtle association between the location of the nostrils on the faces and the personality profiles. During the testing phase, the subjects were presented with more pictures of faces and were asked to judge their personalities based *on their intuition*. Although the subjects did not consciously understand that there was an association between the location of the nostrils and the personality profiles, their judgments of the personalities were based on that rule. What is even more significant about these findings was the fact that the subjects' learned association between the location of nostrils and the personality profiles remained even though they did not receive any feedback from the experimenter

to confirm their belief. In short, information we receive from our own and other people's personal observations and experiences is unreliable and biased.

Authority

We are all continually learning from authority, whether we are watching the news, training for a new job, or reading a textbook. Authorities may be better-informed or more knowledgeable sources of information, or, in some cases, they may simply be people who have the power to dictate what others ought to know and learn. As human societies grow increasingly more complex, labor becomes more and more specialized, forcing people to develop a more narrowly defined set of skills or to specialize in one narrow area. Not only does our society have many different professions, but each profession is also divided into sub-disciplines. Take, for example, the legal profession, wherein a lawyer usually specializes in a specific type of law: bankruptcy, civil rights, criminal, or corporate law. A psychologist usually concentrates his/her research efforts in a specific and narrow area, such as cognition, child development, psychopathology, or neuropsychology. A child developmental psychologist will most likely narrow his/her focus to one specific developmental stage, such as infancy, preschool or adolescence. And as we become more specialized in one area, we become more dependent on the knowledge and expertise of other people (i.e., authority) in the other areas of our lives. Thus, learning from authority has many benefits. It is efficient since one person doesn't have to learn all existing knowledge but can instead master only one set of skills or one profession. And collectively, as a species, we can preserve and continue to develop new skills and to generate new knowledge.

However, a problem may arise if the authority becomes untrustworthy or isn't the expert he/she claims to be. Even the most intelligent people of their time have been wrong about something. Take for example, Aristotle, one of the giants of ancient Greek philosophy, who was once called "the last human to know everything that was knowable during his lifetime" (Hergenhahn, 2001, p. 41). Aristotle was wrong about many things. He erroneously believed that Earth was at the center of the universe and that the stars and the planets were perfect spheres moving in circular motion. However, because he was so greatly regarded, his view of the universe remained unchallenged for a long time.

Reasoning

We can also use rational reasoning to infer what's unknown. Recall from Chapter 1 that the Ancient Greeks, for example Plato, believed that true knowledge could come only from logical reasoning. And although we no longer accept the view that logical reasoning is the ultimate source of knowledge, using this kind of reasoning to make inferences can be quite useful and efficient.

For example, if you know that fire burns, you can infer that touching a hot stove will probably burn you as well; you don't have to actually touch the stove to know what will happen.

This form of knowledge, however, also has its weaknesses. First, an assumption that may seem rational can still be false. For example, it seems rational to assume that as the number of people who witness someone in need of help increases, the number of people who will respond to that need would also increase. However, research has shown that quite the opposite is true. As it turns out, people tend to be less inclined or even slower to help if they think they are not the only witness. This phenomenon is called the *bystander effect*.

Secondly, we may believe that we operate on rational thinking when, in reality, our thought processes are flawed. For example, researchers from Stanford University, Bastardi, Uhlmann, and Ross (2011) conducted a study where people were asked to evaluate two fictitious studies on the merits of their scientific design and to rate the validity of both studies' conclusions. The fictitious studies were measuring the effects of daycare and home care on children's development. Prior to the experiment, the participants filled out a questionnaire in which they revealed their personal opinions about daycare and home care and also the likelihood of their using one or the other for their own children. The researchers wanted to see whether participants' evaluations of the fictitious studies would be based on the actual merits of the studies (i.e., unbiased), on their original personal opinion (biased), or on what they desired to be true (i.e., biased). What the researchers found was that participants were largely biased in their judgments, and that these biases came from their wishful thinking. Specifically, participants who were planning to use daycare services for their own children were more likely to provide evaluation that favored daycare over home care, even if their prior personal opinion favored home-based services. This and many other studies show that wishful thinking and other sources of biases often cloud our reasoning.

2.2 The Birth of Modern Science

So far we have considered three non-scientific forms of acquiring knowledge, and I have tried to convince you that each one of them was flawed in some way. Later in this and the proceeding chapters, you will come to understand why and how the scientific method, if done right, addresses the weaknesses of all three forms of inquiry and yields the most reliable information. But first, let's turn our attention to the distinct time and people who led the Scientific Revolution.

Renaissance

Between the 1400s and 1650s, Europe, and Italy in particular, was experiencing an economic growth and revival after being intellectually repressed during the Middle Ages. At the same time, the church was losing its power. Corruption within the church was one likely reason for its decline, but so too was the fact that people began to show more interest in the natural world and real life

experiences as opposed to being preoccupied with the afterlife. This period in human history came to be appropriately known as the Renaissance, or the "rebirth. The dynamic political and social life of Renaissance Italy, as well as the weakening of the church's authority, gave rise to ***humanism***, an intellectual movement that revived the classical writings, including the teachings of the ancient Greek philosophers, and celebrated the realism of human experiences. While in the art world, this manifested itself in a shift towards capturing more secular and realistic themes, scholars and writers were looking for new and objective ways to study their natural world.

Discovery of Discovery

The notion that science and discovery go hand in hand will probably surprise no one. In fact, we've come to expect breakthroughs and innovations through scientific activities. But, according to historian David Wootton, the concept of discovering something new didn't exist until sometime after the "discovery" of America in 1492. And the reason people didn't believe in discoveries (in the same way we do today) up to that point was the prevailing attitude that everything that was to be known was already known, and simply needed to be uncovered because they were lost or forgotten. In his most recently published book, *The Invention of Science*, Wootton (2015, para. 2) asserts that "Before discovery history was assumed to repeat itself and tradition to provide a reliable guide to the future; and the greatest achievements of civilization were believed to lie not in the present or the future but in the past, in ancient Greece and classical Rome." European discovery of the new land of America, including the new flora and fauna, transformed "how people [Renaissance thinkers] understood their own actions" (Wootton, 2015, para. 2); old beliefs, facts, and theories could be reevaluated and even replaced.

The Birth of Modern Philosophy

Science is the superior form of obtaining reliable knowledge. But what is knowledge and how does it originate? These are important philosophical questions that we need to grapple with if we want to claim that one method of obtaining knowledge is better or more reliable than others. This branch of philosophy came to be known as *epistemology*, deriving its name from two Greek words, *episteme*, meaning "knowledge," and *ology*, *meaning "*science of." Aristotle and other ancient Greek philosophers were the first in the Western world to question the nature and the origin of knowledge. During the Renaissance, two epistemological schools of thought emerged, *rationalism* and *empiricism. These* marked the birth of modern philosophy and played the key role in the emergence of science.

Rationalism vs. Empiricism

René Descartes (1596–1650) led the rationalism school of thought, and his approach marks the birth of modern philosophy. He found the teachings of the ancient Greek philosophers too unsettling—they offered contradictory opinions on the same matters, and everything was up for

debate. Contrarily for Descartes, mathematics was a method by which one could arrive at some exact knowledge by way of deduction from a set of known principles (e.g., axioms) or intuitive facts. Because Descartes believed that the information that had been taught to him was unreliable, he decided to reject all prior assumptions and start from scratch. He proceeded to question the "truths" of everything that he had been taught, including the teachings of the ancient Greeks and his own personal, sensory experiences. He accepted nothing unless there was no doubt in his mind about its certainty. The purpose of such systematic doubting was to establish solid and irrefutable principles (i.e., innate knowledge). This, he believed, could be used as the foundation to answer more complex questions; as in mathematics, where a set of true statements or assumptions would lead you to the exact answer.

Descartes' philosophical approach came to be known as *rational* because it was based on rational inferences for gaining further knowledge. What supported this approach was Descartes' belief that everything was connected, and just like a domino effect, the knowledge of the simplest thing could lead one to uncover the answer to the most complex phenomenon. Notice that rationalism is not an entirely new philosophical movement. The idea of distrusting one's senses and relying on rational thought can be traced back to Plato. We can also draw a parallel between Plato and Descartes' methods of ***deductive*** reasoning, also known as a "top-down" approach of inferring from something general (e.g., an innate knowledge or a pure form) something specific (see Figure 1).

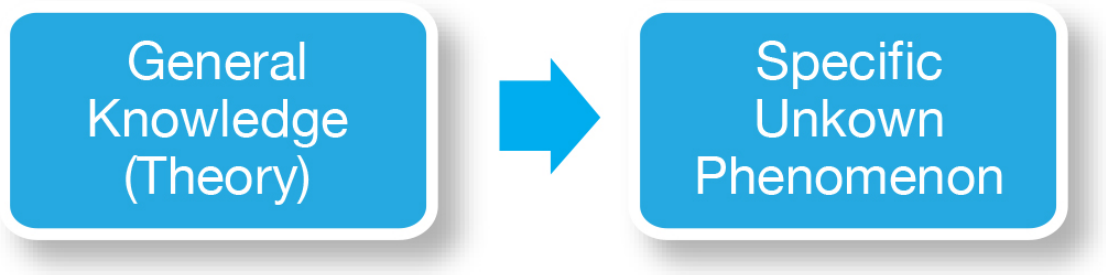


Figure 1. *Deductive Reasoning.*

Another important philosophical figure is David Hume (1711–1776), who is considered by many to be the father of *empiricism*. He distinguished two types of information, or perceptions of our mind: information that comes from our senses and information that is the product of our thought. They differ in degree of "force and vivacity" (Hume, 1748, section 4). Hue felt that the less forceful our thoughts or ideas are, the more forceful are our sensory impressions. Unlike Descartes, who believed that some ideas are innate and should be used to deduce what we don't know, Hume argued that we derive ideas from our observations and memory, from direct sensory experiences of the external world. Hume argued that we employ ***inductive reasoning***, also known as a *bottom-*

up approach, when making inferences about something general and unknown from a specific observation or memory (see Figure 2).

The practice of explaining or predicting things from the evidence of our memory and senses, according to Hume, is based on our human tendency to impose causal relationships between events wherein every event has a prior cause and every cause has a subsequent effect. Inductive reasoning is based on the assumption of the causal link between the present and the past; we can infer one fact from another because we assume that they are causally related. One of Hume's (1748) examples of using induction to make inferences is the following scenario: A man finds a watch on a desert island. This discovery leads him to infer that a person or a group of people previously visited the island, because a watch doesn't naturally exist in a desert island, and someone must have left it there in order for it to be found (section 22).

Another important proposition Hume makes is that knowledge about the connection between things or events does not come to us by way of analytical reasoning, as Descartes argues, but through prior experiences. For example, the conclusion that the desert island must have been visited by at least one other human is not made by way of some analytical reasoning but from our prior knowledge, i.e., experiences, about wrist watches, and the fact that one had to have been worn by a person to appear on a desert island (e.g., as opposed to be naturally existing there). Here is another example he makes in his 1748's essay, *Enquiry Concerning Human Understanding*, where he further justifies his conviction that one's experience is the primary source of knowledge:

Present two smooth pieces of marble to a man who has no tincture of natural philosophy; he will never discover that they will adhere together in such a manner as to require great force to separate them in a direct line, while they make so small a resistance to a lateral pressure. Such events, as bear little analogy to the common course of nature, are also readily confessed to be known only by experience; nor does any man imagine that the explosion of gunpowder, or the attraction of a loadstone, could ever be discovered by arguments a priori. (section 4)

Priori arguments refer to analytical reasoning, or what Descartes calls intuition. Thus, Hume asserts that all laws of nature are discoverable by experiences and not reason.

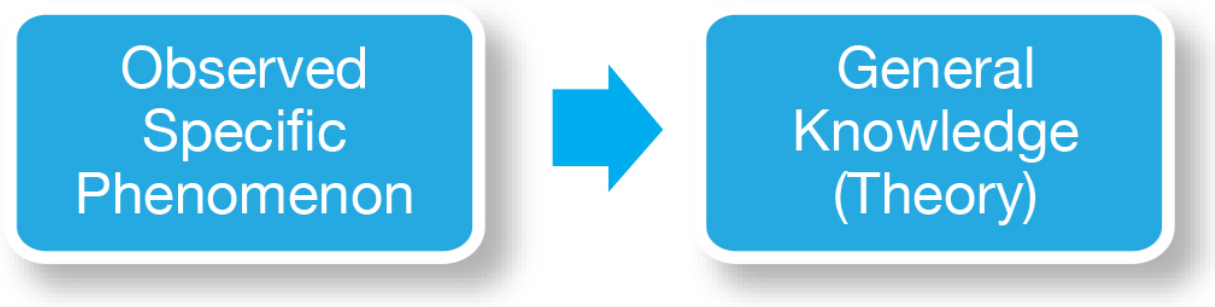


Figure 2. *Inductive Reasoning*

Galileo

Before the discovery of the heliocentric (Sun-centered) system, the view of the universe in Europe was based on Aristotle's model of the Earth-centered cosmos, which was further improved upon by Claudius Ptolemy (c. AD 90–c. 168). In this model, the celestial objects were supposedly attached to spheres, rotating around Earth at different velocities. To account for the varying brightness of the stars, which would indicate that the distance between the stars and Earth had changed, Ptolemy revised Aristotle's model to suggest that the planets were attached not to the spheres but to the circles that he called "Epicycles," and these epicycles were attached to spheres that he called "Deferents." The planets would revolve in a uniform circle on epicycles, while epicycles would revolve while attached to the deferents.

The Earth-centered model fit perfectly with Christian theology, in which humankind, created by God, had to have a central place in the universe. So, this model was embraced and adopted by the Catholic Church as the literal representation of the universe.

Aristarchus of Samos (310 BC–230 BC) earlier suggested that the Sun was at the center of the universe. However, it wasn't until the Renaissance, and thanks to the tenacious work of Copernicus (1473–1543), Tycho Brahe (1546–1601), Kepler (1571–1630), and Galileo (1564–1642), that this once radical view of the universe gained popularity and acceptance. Galileo's role is particularly notable—he used ***empirical observations*** to provide support to the Copernican, heliocentric model. Specifically, Galileo took advantage of a Dutch invention that made objects appear closer than they actually were and built his own apparatus. Later, he improved upon the device and named it the telescope.

By directly studying the planets, the Sun, and the Moon through his telescope, he made discoveries that contradicted the Church's dogma and supported the heliocentric view of the universe. For example, Galileo found that the Moon's surface was similar to Earth's, with mountains and valleys. This, along with other discoveries, called into serious question the Aristotelian division of the universe into a sublunary world—the Earth and everything up to the Moon (made of earth, water, air and fire), and the superlunar world—the Moon and everything beyond that (the unchanging and perfect domain consisted of a single, fifth element called *quintessence*). This proved that the Moon was just like the Earth and made of the same substances.

Perhaps most damaging to the geocentric model was the discovery that the planet Venus showed phases like the Moon, which proved that Venus was revolving around the Sun rather than Earth.

For details and demonstrations on why this discovery was the proof of the heliocentric theory, go to this page on PBS's website:

http://www.pbs.org/wgbh/nova/ancient/galileo-battle-for-the-heavens.html.

2.3 Science as a Form of Inquiry

Basic Stages of a Research Process

Science is a special process of gathering, analyzing, and making inferences about data. It is an attempt to describe, to explain, and to predict events about which we are interested in learning. Although different scientific disciplines may vary in their techniques and certain details, the process can be roughly broken down into the following, overlapping stages (see also Figure 3): ***identifying a problem/generating research questions, formulating a hypothesis, observing and analyzing data, making conclusions/reporting the results, and refining/revising general knowledge/generating more questions.***

When done properly, this process is able to render reliable answers to our questions. However, no knowledge ever reaches the status of an ultimate truth or fact; it only becomes refined, modified, or completely abandoned, depending on the empirical feedback it receives; it is self-correcting. The cyclical (as opposed to linear) nature of this scientific process constantly updates old knowledge and generates new knowledge and more questions. In the following chapters we will trace the stages of a typical research process and discuss each stage in more details.

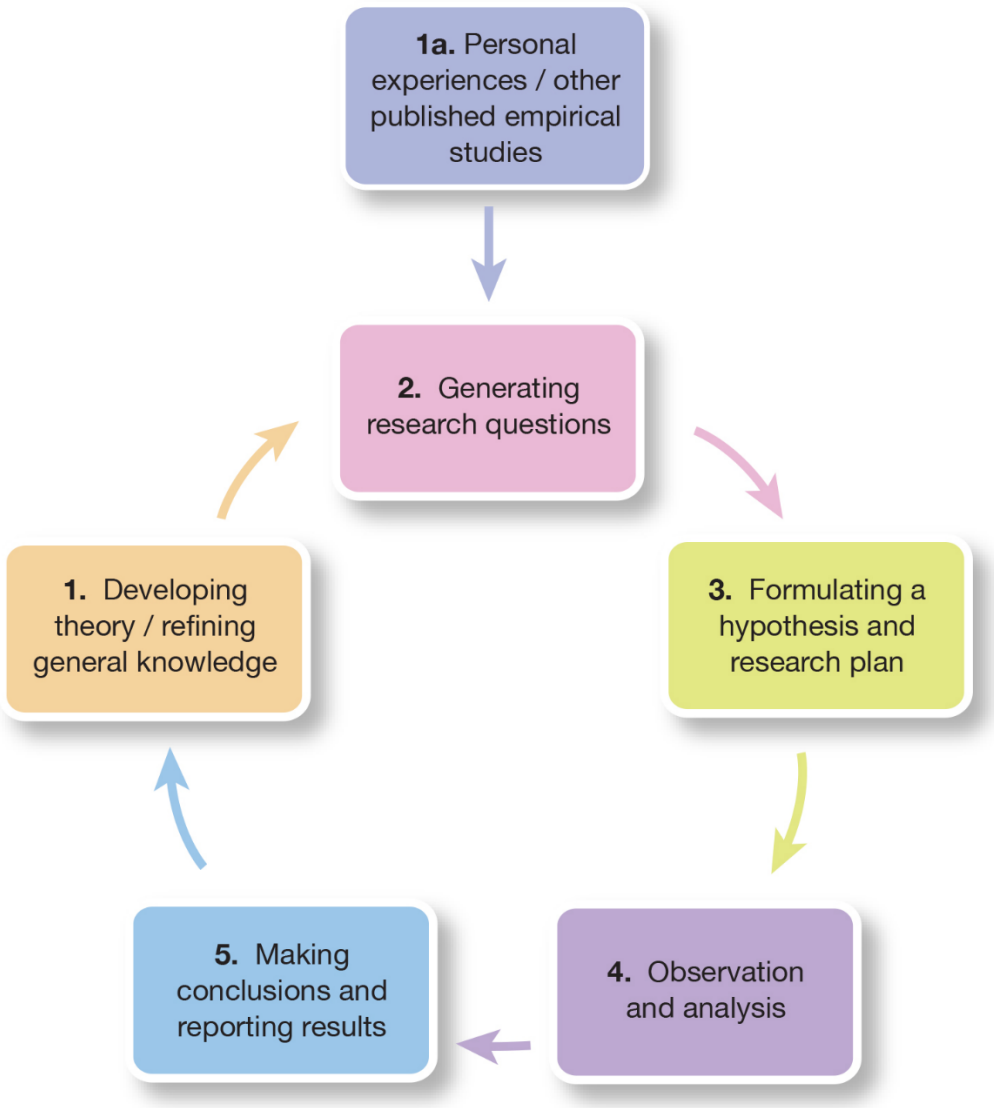


Figure 3. *A Simplified Diagram of the Scientific Process.*

The Key Philosophical Principles of Modern Science

The process of science is not arbitrary and rests on several philosophical assumptions about the nature and the processes of science. The branch of philosophy that deals with these types of questions is known as the philosophy of science. Let's look at some of the key philosophical principles that undergird the scientific process.

Empiricism

Modern science is grounded by observation. However, the observations must be more than casual—they must be *systematic* (i.e., using systematic procedures) and goal-oriented. In other

words, researchers do not mindlessly observe their environment, but rather follow a designed protocol developed around a specific research question based on what they want to know and how they want to select their observations. The ultimate goal is to use observation not only to answer a particular research question but also to make broader contributions to our understanding of the natural world. Galileo is often called the *father of modern science* because his work exemplifies the role of empiricism in science. He used direct observations of the planets, the Sun, and the Moon to prove the Copernican heliocentric theory and to make a larger point about the structure of the universe.

Based on the amount of empirical data and the strength of supporting evidence, the general knowledge derived from observations can be further developed into a *theory*, a statement about a set of concepts which organizes observations and provide an explanation for how and why they are related. Psychological research methods differ in how they make systematic observations and gather their data. Depending on the nature of the phenomenon, some observations have to be made indirectly, i.e. not through researchers' own senses, but by using other tools, like thermometers or an MRI machine. Other observations are done through experimentations, and still others are accomplished by asking questions via surveys, questionnaires, etc.

Determinism

Empiricism can be justified only if observed events have predictable underlying causes as opposed to occurring at random. This important scientific principle is known as determinism. Let's use an example to better understand this idea and how it is connected to empiricism. If you recall, one of Galileo's proofs of the heliocentric system of the universe rests on the discovery of the phases of Venus; sometimes it appears as a dark disk, and sometimes it takes the shape of a crescent. In accordance with determinism, these phases are not random occurrences but have an underlying cause or an explanation. Galileo explained their appearances as the following: Venus revolves around the Sun, which causes us to see it either as a dark round disk or as a crescent (**the illustration of this phenomenon can be seen at:** http://www.telescope1609.com/Galileo.htm). Most importantly, these phases would not be possible in the geocentric system, which thus disproves the Earth-centered, and supports, the Sun-centered universe theory.

Here is another example. Let's assume that a researcher wants to study gender differences in the portrayal of men and women in working roles in TV commercials. It would be a meaningless study unless the researcher believes that there is an underlying cause explaining these differences. For example, he/she can theorize that cultural beliefs and norms about what roles women and men should play in a given society influence how both genders are portrayed in TV commercials, e.g., more men are portrayed as professionals and more women are portrayed as homemakers.

Inductive & Deductive Reasoning

Henry Poincare, one of the greatest philosophers, mathematicians, and scientists, once said, "Science is built up with fact, as a house is with stone. But a collection of fact is no more a science than a heap of stones is a house." What he means is that, although empiricism is a great tool for data collection, it is useless unless we can make sense of what the data mean—we want to connect the dots (i.e., facts gathered from isolated studies) into a coherent theory about our natural world and us in it, in order to better understand and predict future events. To help us understand what we observe, we use both deductive and inductive reasoning. Let's recall the difference between the two processes. ***Deduction*** is the inference about something specific from more general knowledge. In science, this refers to the scientific process whereby one draws an idea and formulates a hypothesis for a study from more general knowledge, a theory or a natural law. This is illustrated in Figure 3, steps 1 through 5. Social psychologist, Kurt Lewin, once said that "there is nothing as practical as a good theory" (see, e.g., Lewin, 1943) to emphasize the important role that a good theory plays in generating more ideas and research. Let me give you an example. Suppose a researcher studies emotions. One known theory about emotions was developed by Paul Ekman, which claims that there are several basic ***universal*** emotions that humans evolved to express in order to regulate interpersonal relationships, such as to make friends (e.g., smiling to signal I can be approached by you) or to keep enemies away (e.g., expressing anger to let the person know you can hurt them). The researcher, who studies smiling and guided by Ekman's theory, may predict that people of different cultures will smile more in larger groups in order to promote collaboration within their groups. And if this hypothesis is supported, it adds more support to the theory of basic emotions.

Induction, on the other hand, is a process wherein one makes inferences about something general, such as a natural law or a theory, from something specific. In our diagram, this is the process 1a through 5. Let me give you another example, demonstrating usefulness of inductive reasoning in science. Suppose a study finds that people perceive smiling differently in different cultures. This specific observation can be used to refine our broader understanding of functionality of emotional expressions. Furthermore, it adds more nuances to the theory of basic emotions by Paul Ekman by suggesting that although such a basic emotion as happiness is universally expressed with a smile, it appears to have evolved to have secondary meaning in different cultures. In fact, studies do suggest that that might be the case. For example, a study by Krys and colleagues (2016) surveyed 5216 people from 44 different countries asking them to rate eight people's faces of either smiling or non-smiling expression, on how honest and intelligent they appeared to be. The found that although in US and some other countries smiling individuals were rated as more honest and intelligent, in other countries the effect of smiling was the opposite, smiling people were perceived as less intelligent and less honest. One interesting factor that affected subjects' perception of smiling was the degree of corruption in their country: High corruption was associated with perceiving smiling people as less honest and vice versa.

Let's use another study to demonstrate how the two processes can complement each other. Anderson and Dill (2000) conducted a study in which their research goal was to find out whether violent video games could cause aggression. The idea for this study came from two sources. One was the real-life tragedy that occurred on April 20, 1999, the Columbine High School shooting, when two high-school students in Littleton, Colorado, shot and killed fifteen students (including themselves) and wounded 23 of their high-school peers. The media at that time reported that the two boys liked to play violent video games and speculated that this behavior could have been the cause of the two shooters' extreme violence. However, the only way we can know if there is, in fact, a causal link between violent video games and aggression is by conducting an empirical investigation, which is what researchers Craig Anderson and Karen Dill did. This is an example of the inductive process. The idea for the study came from witnessing a real-life incident (i.e., the Columbine tragedy) and the results of this study can not only provide us with the answer to the research question but also can help us understand the nature and possible causes of aggression (see Figure 3).

The second source for Anderson and Dill (2000)'s research question comes from the General Aggression Model, which is a theoretical explanation to violent behavior. According to this model, both individual and situational factors can affect people's aggressive tendencies by influencing how an individual perceives and interprets incoming information. A more hostile attribution and perception of other people's intentions can, in turn, elevate emotional arousal levels and lead to more aggressive behavior. This model generates four specific hypotheses tested in Anderson and Dill's (2000) study: (1) long-term exposure to violence is positively related to aggression in real-life situations, (2) short-term exposure to violent video games will contribute to aggressive behavior, (3) individuals with more aggressive personalities will behave more aggressively when provoked, and (4) short-term exposure to violent games will increase the level of aggressive thoughts, which will also be related to aggressive behavior. This is an example of the deductive scientific process, whereby researchers test four hypotheses generated from a conceptual model.

Results of this study can serve to support or refute the General Aggression Model and help us better understand the nature and the causes of aggression. Notice that both the deductive and inductive processes have led Anderson and Dill to the same research questions and to a better understanding about the nature and the correlates of aggression. One final note—it is not required to design a study using both processes. Some studies are born out of a desire to test a specific theory (i.e., deductive process), others—to simply describe a phenomenon (i.e., inductive), and still others may fall somewhere in between.

Falsifiability

An important criterion that separates science from non-science is the principle of falsification, proposed by philosopher Karl Popper. Under this principle, a statement or a theory is considered to have scientific status if it is empirically refutable, i.e., can be disconfirmed. Recall that

a ***theory*** is a statement about a set of concepts which organizes observations and provides a coherent explanation for how and why they are related. A truly scientific theory provides enough details to specify the conditions under which events of interest have to happen, which, in turn, tells us about the events that cannot happen when necessary conditions are not met. Popper calls these predictions "risky" because they run the risk of being wrong—these predictions are so specific that when they fail to be observed, they serve as evidence against the theory they were intended to support. And since it is almost always possible to find confirmation for a theory, only predictions that could be disconfirmed can elevate a theory to scientific status.

Support for a theory that comes from failing to disconfirm its prediction is harder to provide, but it is much stronger and more valuable than support received by confirmation. As Popper (1963) stated, "Irrefutability is not a virtue of a theory (as people often think) but a vice" (p. 36). At least two factors can undermine the scientific status of a theory. One factor is when a theory's predictions are so broad or vague that almost any conditions or observations fall within its prediction realm and virtually no observations can refute it. Let's use Astrology to illustrate this point. Astrology can be defined as the study of how heavenly bodies—the Sun, Moon, and the planets—influence the events and the people on Earth. Here is a segment of a prediction of the events and circumstances that a specific individual could expect to encounter, as reported by the Café Astrology website:

During this period you are motivated by the desire to be of service to others. You may work in some field related to the healing professions or the service industries. Job satisfaction is wished for, but may be hard to realize now, because the ideal may contrast with the reality of the work situation. Care may be needed with respect to the effects of drugs and alcohol on health.

Sample Year of Transits Reports, Before Dec 15 2013–Beyond Dec 31 2014, Transiting Neptune is passing through your 6th House.

The circumstances and the events in this forecast are so broadly defined that almost any future events can fall within the range of this description. These predictions cannot be falsified—you could never prove that such events never happened, only that they did happen.

The second factor is ***testability***. Ideas proposed in a theory have to be measurable before they can be refuted. Freud's psychoanalytic theory is a good example of a theory that does not meet this requirement. The main thrust of Freud's theory is the idea that our unconscious forces—id, ego, and superego—shape our personalities and behaviors. No one, not even Freud himself, was ever able to measure these forces, which is not surprising since they are not accessible. Then, the question is, how can these ideas be ever tested and refuted? Freud's psychoanalytic theory has never been disconfirmed because it is simply not testable. All questions about religion or any supernatural existence face the same issue—they are non-testable, and so they fall outside the realm of science.

The *Differential Susceptibility Theory* is an example of a psychological theory that satisfies the principle of the falsifiability (DST; Belsky, 1997a, 1997b, 2005; Belsky, Bakermans-Kranenburg, & van IJzendoorn, 2007). According to this theory, individuals who are neurobiologically more susceptible to environmental influences will be affected more positively than other individuals when environmental impact is of a positive nature (e.g., nurturing parenting) or more negatively when it is of a negative nature (e.g., stress). Neurobiological susceptibility has been measured at multiple levels—behavioral, genetic, neuroendocrine, and epigenetic. The predictions are also very specific and can be easily refuted if they are wrong. For example, to refute the theory, one would have to show that children with neurobiological susceptibility would show no differences (i.e., no benefit and no vulnerability) in their development to children with low environmental susceptibility under both positive **and** negative conditions.

One final point needs to be made, and that is that the principle of falsifiability does not place any judgment on a theory's importance or potential truth (Karl Popper, 1963, p. 63). It is quite possible that a theory that currently is not testable can someday be tested and supported by empirical evidence, or that the predictions can be made more precise to make a theory falsifiable.

Test Your Understanding of What Constitutes Science:

Is Egyptology a science? Use the key scientific principles to justify your answer.

Here are some facts about this field that should help you answer this question:

Egyptology is the study of the history and culture of ancient Egypt.

One of the questions that Egyptologists are still trying to unravel is how ancient Egyptians built their pyramids. So much physical effort and knowledge had to go into building these marvelous structures that some have even speculated that the ancient Egyptians were in contact with aliens or that a lost and more sophisticated civilization, Atlantis, built them. All explanations remain only myths. Egyptologists don't have any doubts that the Egyptians did, in fact, build them, and several working theories have been proposed to explain how they managed to construct the pyramids without the use of all the advanced technologies we have today. To test one of them, in 1950, the museum of Science in Boston, and a model-maker, Theodore B. Pitman, constructed a model of the Giza pyramids, replicating the hypothesized methods of Egyptian engineering (see Dunham, 1956). Click here to access the article.

Go to the PBS website page at:

http://www.pbs.org/wgbh/nova/ancient/who-built-the-pyramids.html for additional information about current knowledge of who build the pyramids and how they were built.

References

Anderson, C. A., & Dill, K. E. (2000). Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. Journal of Personality and Social Psychology, 78, 772–790.

Bastardi, A., Uhlmann, E., L., & Ross, L. (2011). Wishful thinking: Belief, desire, and the motivated evaluation of scientific evidence. Psychological Science, 22(6), 731–732.

Bakermans-Kranenburg, M. J., & van IJzendoorn, M. H. (2007). Genetic vulnerability or differential susceptibility in child development: The case of attachment [Research review]. Journal of Child Psychology and Psychiatry, 48, 1160–1173.

Belsky, J. (1997a). Variation in susceptibility to rearing influences: An evolutionary argument. Psychological Inquiry, 8, 182–186.

Belsky, J. (1997b). Theory testing, effect-size evaluation, and differential susceptibility to rearing influence: The case of mothering and attachment. Child Development, 68, 598–600.

Belsky, J. (2005). Differential susceptibility to rearing influences: An evolutionary hypothesis and some evidence. In B. Ellis & D. Bjorklund (Eds.), Origins of the social mind: Evolutionary psychology and child development (pp. 139–163). New York: Guilford Press.

Descartes, R. (1996). Discourse on method and mediations on first philosophy. D. Weissman (Ed.). Yale University. (Original work 1596–1650).

Dunham, D. (1956). Building an Egyptian pyramid. Archaeology, 9, 159–165.

Fisher, R. A. (1950). Statistical methods for research workers. London: Oliver and Boyd. p. 80.

Hergenhahn, B. R. (2001). An introduction to the history of psychology. Wadsworth Thomas Learning.

Hill, T., Lewicki, P., Czyzewska, M., & Schuller, G. (1990). The role of learned inferential encoding rules in the perception of faces: Effects of nonconscious self-perpetuation of a bias. Journal of Experimental Social Psychology, 26, 350–371.

Hume, D. (1748). An enquiry concerning human understanding. Retrieved from http://darkwing.uoregon.edu/~rbear/ren.htm

Krys, K., Vauclair, M., Capaldi, C. C., Lun, V. M., Bond, M. H., Dominguez-Espinosa, A....Yu, A. A. (2016). Be careful where you smile: Culture shapes judgments of intelligence and honesty of smiling individuals. Journal of Nonverbal Behavior, 40, 101-116.

Popper, K. (1963). Science as falsification. Conjectures and Refutations. Retrieved from http://faculty.washington.edu/lynnhank/Popper-1.pdf

Sample Year of Transits Report (n.d.). Retrieved July 30, 2014 from Café Astrology.com: http://www.cafeastrology.com/sampleyearoftransits.html

Wootton, D. (2015). The invention of science: A new history of the scientific revolution (EPub). Penguin Random House UK.

Chapter 3: Thinking Like a Scientist

Introduction

The purpose of this chapter is to walk you through the general steps involved in thinking and planning phases of a research project—*identifying a problem, generating research questions, and formulating falsifiable hypotheses*. All three steps will require a thorough review of the literature to be informed on the subject matter. Specifically, before a researcher can launch a study, he/she must be fully aware of the questions that have already been addressed and what is yet to be understood about the phenomenon of interest. Finally, a research plan will have to specify a research design, variables, and their measures to properly collect the data.

3.1 Knowing the Extent of One's Ignorance

"Real knowledge is to know the extent of one's ignorance." — Confucius

It's an interesting paradox that I think many of my colleagues will be able to relate to: The more you learn about a particular subject area, the more you feel like you have more questions than answers. And vice versa, the less you know, the more you feel confident in your knowledge.

Charles Darwin (1871) wrote that "ignorance more frequently begets confidence than does knowledge" (p. 3). He was right. Studies have shown that most people tend to overestimate their abilities or attributes. For example, people tend to believe they are smarter, more honest, or even more attractive than the average person. Social psychologists call it the "better-than-average phenomenon." A study by Kennedy, Lawton, and Plumlee (2002) found that students who performed poorly on exams were more likely to overestimate their future exam grades. On the other hand, better-performing students underestimated their performance and their expected exam grades. The authors explained this phenomenon the following way: "If students are unaware of their poor performance on tests, they are unlikely to realize their limitations… They don't know what they don't know."

The good news is that the same students became better at recognizing their limitations over time, presumably after receiving more feedback from tests and their professors. Essentially, they became more informed in the subject area and aware of the limits of their own knowledge, and they were able to judge more objectively the discrepancy between the two.

An expert or a scientist, on the other hand, is usually keenly aware of his/her limits of understanding – just try asking them a question. You will rarely get a straight answer. Sam Harris, a neuroscientist and the author of "The Moral Landscape," gave a great example of what he calls being a modest expert:

> *"In my experience, arrogance is about as common at a scientific conference as nudity. At any scientific meeting you will find presenter after presenter couching his or her remarks with caveats and apologies. When asked to comment on something that lies to either side of the very knife edge of their special expertise, even Nobel laureates will say things like, 'Well, this isn't really my area, but I would suspect that X is…' or 'I'm sure there are several people in this room who know more about this than I do, but as far as I know, X is…' " (p. 124)*

This modesty comes from the realization that things are nuanced and complicated, and that our understanding of them will always be restricted. Some of those limitations come from the imperfections of our own mind, for example, from our biases. Other times, these limitations are caused by the imperfections of the research methods and by simple human errors. And of course, limitations are borne out of the constraints that scientific principles naturally impose on researchers to safeguard against biases, errors, and other pitfalls. So let's go through a typical scientific process and identify important elements that place constrains but that also safeguard research to ensure reliability.

3.2 Stating a Research Question and Specifying a Hypothesis

Steps 1 & 2

Let's recall the most basic steps usually taken in conducting a research study. Figure 1 should be familiar to you from chapter 2. Keep in mind that because this textbook is intended for beginners, the described process is simplified and doesn't account for all the possible variations in real life research.



Figure 1. *Basic Steps in Research Process.*

Usually, a researcher will conduct a study to test a novel idea, e.g., a theory that provides an explanation for the existence of societal stereotypes. Other times, the purpose of research is to address ambiguity or inconsistencies in existing knowledge on a subject matter, or to test competing theories to add more clarity to the understanding of a phenomenon in question. For example, a researcher may seek to find *empirical* evidence for the fact that infants are more cognitively advanced than is posited by the Piaget's theory of cognitive development. Generating a good set of ***research questions*** can be considered the starting point in the research process because it provides a researcher with the motivation and the purpose for conducting a study. Furthermore, these research questions will set the tone for the entire project, providing the focus and the boundaries. To underscore the importance of having a good research question, Albert Einstein once (allegedly) said, "If I had an hour to solve a problem and my life depended on the solution, I would spend the first 55 minutes determining the proper question to ask, for once I know the proper question, I could solve the problem in less than five minutes."

Characteristics of a Good Research Question

The following three characteristics define a good research question:

- ***Falsifiable:*** A question must be testable. For example, spirituality and religious questions cannot be addressed or tested and thus, fall outside the realm of science.
- ***Clearly articulated:*** Even the most innovative idea is useless if it cannot be coherently expressed and conveyed to others.

- ***Grounded in previous research:*** It is imperative that any claims or inquiries are made after consideration of all prior scientific work conducted and reported on the subject matter in peer-reviewed journals (i.e., journals that conduct a rigorous review of every submitted article by several experts in the field and that publish only the ones that are approved by the experts); only information found in peer-reviewed journals provides the most reliable information on a topic of interest and only *they* should be used to provide scientific justification for your research question. If no prior scientific work can be found on the topic of your research question, ask yourself why. Does it belong to the realm of science? Perhaps the questions you're raising can never or cannot *currently* (given the absence of proper measuring instruments) be addressed *scientifically*. Alternatively, is it possible that the question isn't significantly important enough or the answer is too obvious to be worth one's time and effort investigating? This brings us to the last important characteristic of a good research question.
- ***Carries important implication:*** The literature review should help you see the relevance and the importance of your research question. Remember that when you communicate your results to the public, either through a poster presentation or by publishing your results in a journal, you will have to convince your reader of the importance of your study and its contribution to the larger body of knowledge.

Characteristics of a Scientific Hypothesis and a Theory

The next step is formulating a ***scientific hypothesis***, which is a statement that makes a specific prediction about a phenomenon of interest. This prediction should be about a presence rather than an absence of an effect. For example, a researcher can predict and subsequently test the presence of some group difference (e.g., gender difference in communication) but a researcher cannot predict and test a hypothesis that the group difference doesn't exist. To give you an example, I may hypothesize that playing violent video games causes aggression; the effect to be tested in this hypothesis is aggression. I cannot test the opposite, that playing violent video games does not cause aggression--the absence of the effect. The explanation for such a limitation in hypothesis testing will require first understanding the nature of the statistical hypothesis testing, which will be the topic of chapter 8. Ideally, a hypothesis is derived from a ***theory,*** which is an explanation of some phenomenon based on substantiated evidence. In other words, an existing theory should generate predictions about what we should expect to observe if this theory is correct.

Furthermore, to be deemed scientific, a hypothesis must be testable. Recall the principle of ***falsifiability*** (falsification) (see chapter 2)—a claim or a theory can gain scientific status only if it can be empirically disconfirmed.

While a disconfirmation strategy is indeed a powerful way to test a hypothesis, confirmation is another widely used method in scientific testing. Let's take a closer look at the both strategies.

A Confirmation Strategy. When a researcher tests a prediction that is supposed to occur under specific circumstances when the claim/theory is correct, he or she is using a confirmation

strategy ***to confirm the claim/theory***. For example, Svanberg, Mennet, and Spieker (2010) conducted an intervention study with the main purpose of promoting the development of secure attachments in young children of families residing in a low-income, urban British community. Secure attachment is a concept that comes from *attachment theory*, developed by Bowlby and Ainsworth (Bretherton, 1992), and is theorized to have a positive influence on children's development. For example, children who are securely attached to their caregivers are more likely to be socially and emotionally well-adjusted, have better academic achievements, and be more compliant when compared to their counterparts.

Attachment theory also suggests that an early maternal parenting style is the cause of differences in attachment styles; children of sensitive mothers are more likely to develop secure attachment and while children of mothers who lack sensitivity are more likely to develop an insecure attachment. Returning to the intervention study, the researchers trained mothers to use more sensitive parenting styles by focusing them on their children's emotional needs. The prediction was that, if attachment theory is correct, more children would develop secure attachments to the mothers who received the training, thereby confirming the theory. The results of the study did, in fact, confirm the attachment theory with the finding that more children had developed secure attachments to those caregivers who practiced more sensitive parenting styles.

We can represent this confirmation hypothesis testing strategy with the following statements:

If the attachment theory (A) is correct,
then B (sensitive parenting leads to secure attachment) should be observed.
Therefore, If B, then A is supported.

A Disconfirmation Strategy. Disconfirmation strategy is defined as when the prediction being tested is *not* what is expected to happen according to a given theory. Disconfirmation is considered a more powerful strategy to provide empirical support than a confirmation strategy because people tend to pay more attention to and look for things that confirm their beliefs and opinions and dismiss things that do not. This is a phenomenon known as *confirmation bias*. Confirmation bias weakens supporting evidence. Since researchers are no exception and can fall victim to confirmation bias, using disconfirmation strategy is a way to keep researchers' own biases in check and still be able to find supporting evidence if it's there. **Furthermore, a disconfirmation strategy is more powerful because it takes only one instance with a negative finding (a finding that does not support a theory) to disprove a theory whereas positive findings, while confirm a theory, will always leave the possibility that it may be false, and the error or the conditions under which the theory is false has not been simply detected.**

How would the same researchers, Svanberg, Mennet, and Spieker (2010), apply a disconfirmation strategy to test the attachment theory? The most obvious, way would be to ask parents to practice

insensitive parenting. Since this would be unethical and immoral, an alternative approach is to administer two different interventions to two different groups of families. One intervention would be to train parents to provide a more sensitive parenting style (the original approach) and the other would be to give the parents psychological counseling without providing education on sensitive parenting techniques. The disconfirmation strategy in this case can be broken down into the following statements:

If the attachment theory (A) is correct,
then C (psychological counseling to caregiver) is NOT more likely to lead to secure attachment than B.
Therefore, If C, then A is supported.

Both, confirmation and disconfirmation strategies are acceptable in science, as long as the hypothesis that is being tested is falsifiable.

One important characteristic of a falsifiable hypothesis is that the stated prediction can be directly or indirectly observed. Direct observations are made through our senses: smelling, seeing, touching, or hearing the phenomenon being predicted. For example, a researcher who makes the claim that playing violent video games causes aggression could make the following prediction: *playing more than 5 hours of violent video games per day is associated with increased physical violence in young children (e.g., hitting, pushing)*. This hypothesis can be easily and directly tested by measuring the length of time children play violent video games and also directly observing their behavior on a playground when they play with other children. The same hypothesis can be also tested by conducting indirect observations. For example, a researcher can measure the amount of time people spent playing violent video games and their subsequent behavior by simply asking questions through a questionnaire.

Just like a good research question, a good hypothesis has to be clearly stated and based on prior research. For example, predicting that at least *5 hours of violent video game play can lead to increased aggressive behavior* must have justification, scientific or logical. Perhaps, previous studies found a link between 5 hours of video game play and physiological changes that can potentially be associated with aggression. Remember that almost any imposed specificity will have to be backed by evidence or good reason. This will save a researcher time and resources because the more evidence that can be found to support a hypothesis, the better the chances that the hypothesis will be supported, and all the time and money spent on a study will not go to waste.

Important Concluding Remarks: The more specificity a theory or a hypothesis provides, the more opportunities it presents to gain scientific support either through confirmation or disconfirmation strategy if, in fact, it is correct.

Why Confirmation Strategies in Social Sciences are More Common?

Despite the fact that a disconfirmation approach is a more powerful way to test a claim, more studies tend to take a confirmation approach. For example, a study by Uchino, Thoman, and Byerly (2010), analyzed studies published in the Journal of Personality and Social Psychology over 23 time period. Their analysis revealed that about 77% of the studies used a confirmation approach in hypotheses testing. So the question is why? One plausible explanation is that most proposed and tested hypotheses are non-absolute, which means that they predict that certain effect or differences will be found some of the time, but not all of the time. Given these non-absolute nature of science, a confirmation strategy is more useful. For example, if a theory predicts that insecure attachment leads to social adjustment sometimes, finding out that some subjects did not develop such issues will not disconfirm the theory. Of course, we can also question the usefulness of such non-absolute theories. But even in other scientific fields, non-absolute theories are quite common. For example, smoking can cause cancer in smokers, but not in all of them (My grandfather smoked a lot but never developed cancer) (i.e., Sanbonmatsu, Posavac, Behrends, Moore, & Uchino, 2015).

The Link Between a Research Question and a Hypothesis

A research question and a hypothesis are directly linked: while the former sets the tone and raises the question that will drive the study, the latter specifies how the question will be addressed by predicting what is supposed to happen. For example, researchers Moll and Meltzoff (2011) studied the topic of visual perspective-taking in 3-year-olds. Visual perspective-taking is part of a larger topic known in developmental psychology as the *theory of mind*, which is an understanding that other people can have beliefs and perspectives, i.e., visual perspectives, desires, and information, that may be different from one's own.

This knowledge appears to develop in children between four and five years of age, when children can explicitly verbalize what another person feels, sees, or thinks in a given circumstance, even if the other person's perspective contradicts that of their own. In other words, they can simultaneously hold two different perspectives and accept the possibility that two different perspectives can coexist simultaneously. However, studies have also shown that understanding of conflicting desires emerges at around three-years of age, which is earlier than children's understanding of others' beliefs and knowledge.

The research question that Moll and Meltzoff (2011) decided to raise was whether children as young as three would also have an understanding of differences in visual perspectives. Based on the extensive review of prior studies on children's perspective-taking, researchers hypothesized that three-year-old children would be likely to pass a visual perspective-taking test if the tasks that a three-year-old child were instructed to do were easier to understand and follow. In other words, the researchers hypothesized that three-year-old children would understand false visual perspective beliefs (just like they understood conflicting desires) if they tested them with an age-

appropriate measuring instrument. Notice that the justification for making such a specific prediction was grounded in prior evidence of children understanding conflicting desires at age 3.

To test their prediction, Moll and Meltzoff (2011) developed a new testing technique, called a *color filter technique*, that would require children to recognize and distinguish colors rather than word pairs (e.g., *left-right* or *in front of-behind*), an ability which may not always be acquired by the age of three. Children would also have to respond to questions by manually pointing to or placing an object rather than verbalizing their answers. In one of the experiments, the children were presented with a pair of blue objects, a pair of dogs. While to the children, on the children's side of the table, the two objects were clearly seen as blue, to an adult, one of the objects appeared green because he/she was looking at one of them through a screen with a yellow filter from the opposite side of the table. Children were given an opportunity to see the adult's perspective.

To test the children's understanding of the adult's different visual perspective, the adult would ask the child to put the "green" dog into a bag for him/her. The child would have to decide which one of the two identical objects the adult was referring to; i.e, the child had to try to take the adult's visual perspective. This test confirmed the prediction and demonstrated that three-year old children were able to take on another person's visual perspective.

THINK LIKE A SCIENTIST AND ASK YOURSELF THESE QUESTIONS:

What is the research question that the hypothesis of your study will address?

What is the important implication that your study might have on the understanding of the issue?

3.3 Literature Review

Some Vocabulary

- ***Scientific knowledge*** is knowledge acquired through science and shared with the public in academic journals, conferences, and books. Before being published in academic journals and books, articles and chapters are reviewed by the experts in their field, i.e., they are ***peer-reviewed***, to make sure that only high-quality studies and reviews are published.
- ***An empirical article*** reports on a scientific study conducted by an author(s) of the article.
- ***A theoretical article*** reports on a theory.
- ***A review article*** provides a brief or an extensive review of an existing body of empirical and/or theoretical work (scientific articles) written on a particular topic.

- ***A primary source*** is a publication of an original study/theory.
- ***A secondary source*** is a publication that reviews/cites a primary source.
- ***A peer-reviewed journal*** is a scholarly journal where the articles are written by experts in their field. They go through a rigorous review by other experts before publication, for quality control.

Purpose of the Literature Review

Science should be progressive in that it should advance our understanding of the world. Therefore, a good theory or idea (which is typically what drives a new research) should be able to explain past findings and predict future ones. The only way a researcher would know that he/she is dealing with a plausible theory is by keeping track of all the studies on the subject matter--i.e., knowing the body of evidence. Thus, reviewing relevant scientific literature is an essential part of any research process because it informs a researcher of what has already been done to address his/her research question and whether there are any gaps in the current understanding of the phenomenon. Thus, one of the reasons for conducting a literary search is to help the researcher build a worthwhile research question; a question that is interesting and important but also hasn't been adequately addressed by previous studies. Another reason for this search is to guide the researcher in the prediction about the outcome of the study. For example, the hypothesis that children as young as three would be able to take on another person's visual perspective was not made arbitrarily by Moll and Meltzoff (2011); it was grounded in previous findings about children's perspective-taking skills.

The role of literature review isn't just limited to assisting researchers in asking the right research questions and formulating sound hypotheses. Literary analysis should inform the entire research process, from establishing research and hypothesis, to choosing the appropriate research design, study's variables, and their measures, to deciding on the statistical analysis and the interpretation of the results. For example, a researcher who wants to look at the link between the playing of violent video games and aggressive behavior should draw upon both the strengths and weaknesses of all prior empirical and theoretical work on the same or similar question. Whatever worked well should be replicated, while whatever didn't must be either revised or avoided altogether.

Finally, after obtaining results, the researcher has to refer back to the literature once again to make the connection between the findings of the present study and all prior knowledge. Do the new findings add to what is already known, or do they contradict it? If it's the latter, the researcher will have to explain why this might be so, refine current knowledge on this topic, and call for further investigations. In a way, this is both the last and the first step of a research process—the last, because it concludes the study, and the first, because it gives rise to more ideas and reasons to conduct additional research (see Figure 1). Thus, research is a circular process in which new studies are constantly updating previous knowledge.

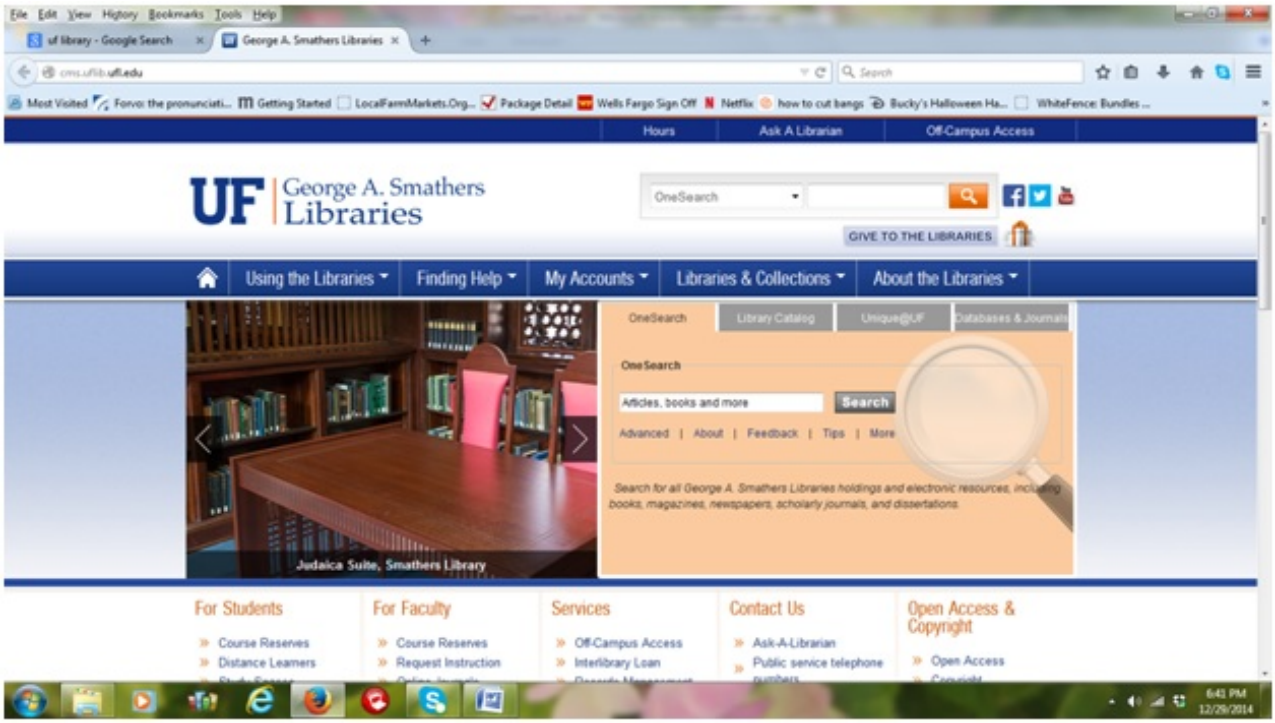
Important Concluding Remarks: A literature review should inform every aspect of a research process: from developing a research question and a hypothesis or choosing a research design, to connecting results of a study to a larger body of existing knowledge.

How to Conduct a Literature Search

Before there was the Internet, there were libraries. Of course, we still have libraries, but many of them now offer access to their books and journals electronically, so you can search and read many articles and books in the comfort of your own home or while sipping on a nice cup of coffee at a coffee shop.

Suppose my research question is whether playing violent video games can cause aggression in children. My first step would be to find out as much as I can about what has already been reported on this topic. To give you an example of how to conduct a literature search, I will use the UF George A. Smathers Library, to which I and many of my readers have access.

1. Begin by Going to UF libraries website and select either ON- or Off-Campus Access (if you are not using UF computers on campus):



2. Connect to UF e-resources by following the instructions on how to connect if you are on or off campus.

Connecting to UF Library Resources from Off Campus

Off-campus access to eResources: restricted to UF students, faculty, and staff
UF Libraries purchase special licenses for electronic items such as eJournals and articles, eBooks, databases, and other digitally available materials, so off-campus access to these materials is limited to the students, faculty, and staff of the University of Florida.

CONNECTING FROM ON CAMPUS?
If you are connecting to the Internet from any UF network, you should be able to connect directly to our eResources.

CONNECTING FROM OFF CAMPUS?
If you are doing research from **OFF campus** (home, coffee shop, etc.) then you must login before you can access any of our electronic materials.

FOR ACCESS, CHOOSE FROM TWO OPTIONS:

OPTION #1:UF PROXY SERVER

- **GREAT** for **FAST** searches.
- NOT ideal for lengthy research sessions (see Option #2)

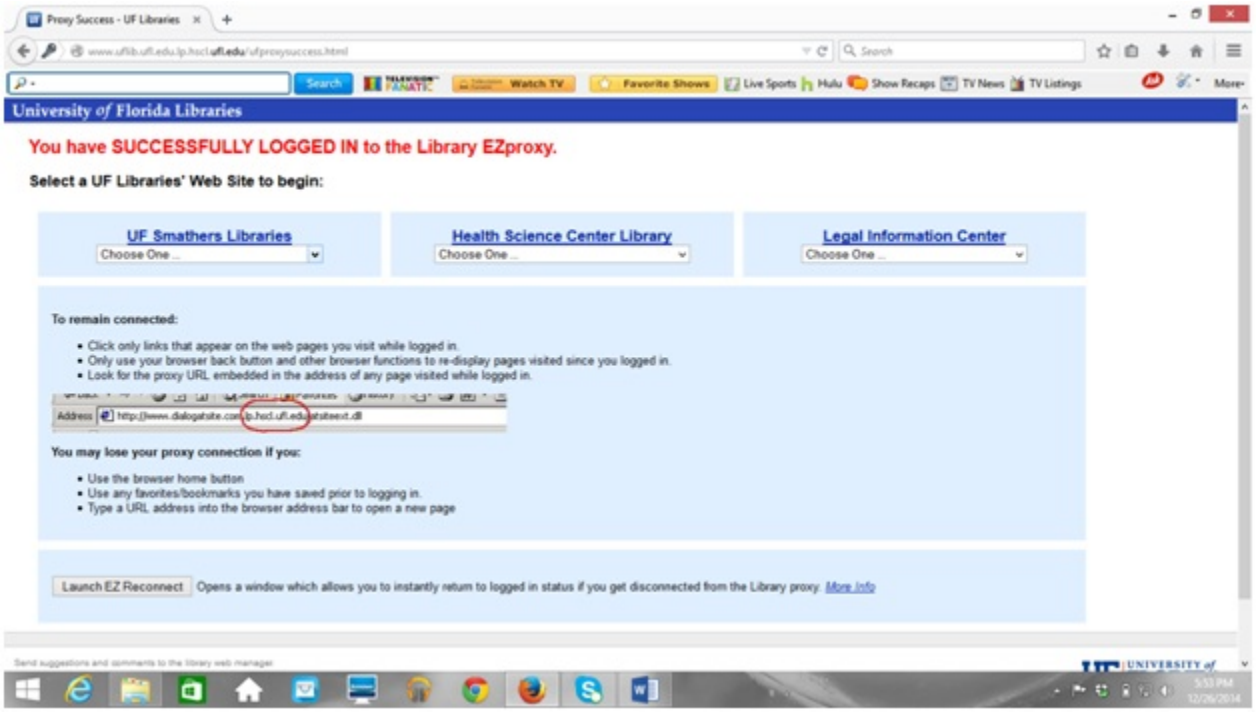
Follow this link to login with the UF proxy server.

Forget your Username and/or Password?
Need to reset your password?
Call: 352-392-HELP

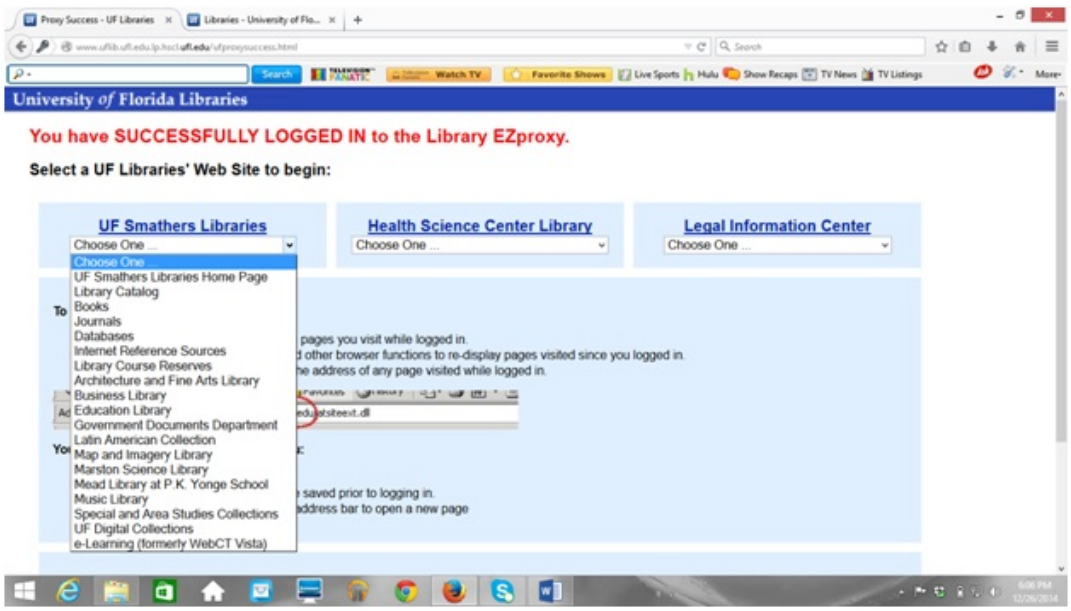
OPTION #2: VIRTUAL PRIVATE NETWORK (VPN)
[Free Software keeps you logged in]

- This is the BEST option for lengthy research sessions
- Free Software
- Free Tech Support Available: 352-392-HELP

3. Once, successfully logged in, you should see the following screen.



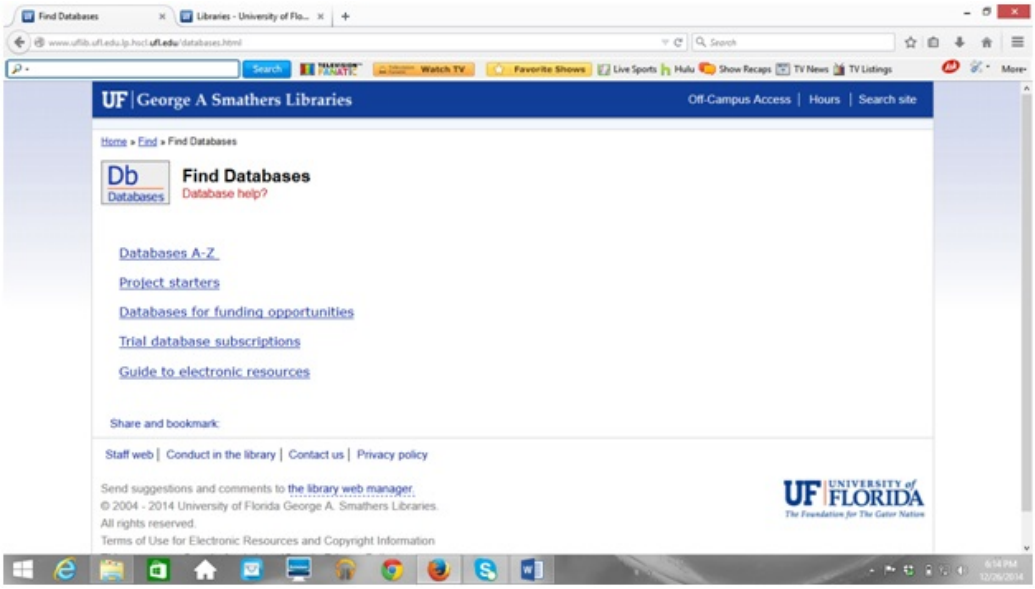
4. Several options are available to start a search.



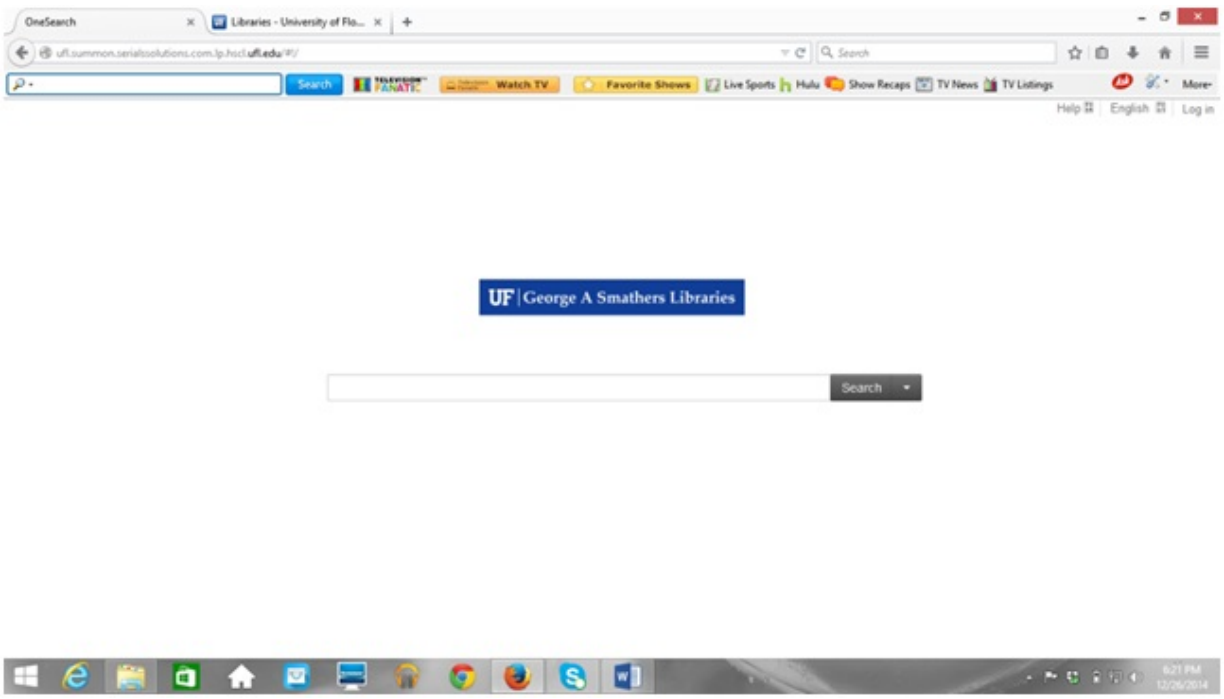
5. Choose options 'Books' and 'Journals' if you have a specific book or a journal in mind. If not, choose 'Databases' option to search all available journals and books on a particular topic.

6. Choose 'Project starters' option to start the search.

It works similarly to a Google scholar search and searches electronic resources subscribed to or purchased by the UF library. So, it will give you full text access to most materials. However, the library grants access to other databases, such as Academic Search Premier and ProQuest, and you can certainly try different databases.



7. Here is what you should see when you choose 'Project starters':



If the topic of your research is new to you, start by typing something general. For example, I can type 'aggression and video games' and see how many different articles I can acquire.

Narrow your search to 'Peer-Review' articles, 'Journal Article,' and 'Psychology' if you get too many results. You can also limit your results to only recent publications by entering the year and the month 'from' and 'to' to specify the range of the publications' dates.

If the search still gives you too many articles, you can refine it by entering a more specific subject term.

The results of my search gave me 824 articles, which is quite a bit. But this also lets me know that my topic had received a lot of scientific attention, and that its body of literature is rather large. One of the things that might be helpful for someone who is reviewing a large body of literature is to find the most recent review article that gives a good summary of all the most pertinent studies on the topic. For example, when I added the term 'review' to my subject term 'aggression and video game,' I found a review article by Anderson et al. (2010), providing a comprehensive review of the studies done on aggression and video game play.

Although a review article, like the one I found in my literature search, gives a good summary of the articles on the topic of one's interest, you should find and read as many primary sources as possible. Many details that may be important for your own study might not be mentioned in a review article.

3.4 Quality of Sources

Professional Journals

Not all journals are equally reliable sources of information. First, you have to look for evidence in professional, peer-reviewed journals. Typically, these journals are published by professional organizations whose goal is to advance knowledge in their respective disciplines. These journals are called "peer-reviewed" because all submitted papers must go through a review process by the professionals in the area of the submitted topic. The more respected the journal is the tougher is the review process. Just to give you an example, here is an excerpt from one of reviewers of a journal that I submitted my manuscript, in hopes to get it published there. As you will see below, the reviewer' assessment was not very favorable, and, my manuscript was rejected. As painful as it is to receive negative feedback, it is meant to help authors improve their skills and produce better research.

> *...I was very excited when reading the paper's abstract - which made me decide to actually review the paper - and there are several reasons for this. Since the advent of video games, carrying out content analysis for video games has always been a challenge and I was particularly interested in how author(s) address this - methodological - challenge, which, I think, is the reason we have so few well-conducted content analysis for video games. Second, author(s) also focused on moral content - a topic which, again, I see as one of the most interesting challenges for researchers. The interactive nature of video games makes them particular well-suited examples for studying the relationship between media and morality. However, and sorry to put it this bluntly, when reading the paper, I was rather disappointed as the paper missed to address these two crucial aspects. Therefore I do not see the paper ready for publication and I recommend author(s) to continue their research - which in itself is very worthwhile - but to use this study as a pre-study for a much more thoroughly conducted main study that is carried out with more methodological rigor and a deeper theoretical interrogation...*

Ouch, still hurts to read it!

One, reputable institution, for example, is the American Psychological Association (APA). Its mission is "*to promote the advancement, communication, and application of psychological science and knowledge to benefit society and improve lives*." The APA publishes numerous journals on topics ranging from psychology of arts and aesthetics (*Psychology of Aesthetics, Creativity, and the Arts*) to neuroscience and brain development (*Behavioral Neuroscience*).

Normally, professional organizations care about their reputation and the quality of their work. Thus, these journals will try publishing only high quality papers. Of course, the quality of these journals still varies: There are great, top-notch journals, and there are some journals that are just

okay. The most common strategy to distinguish low quality journals from high quality journals is by their ***impact factor***. Calculated by Clarivate Analytics, impact factor is "*the average number of times articles from a two-year time frame have been cited in a given year*" (www.apa.org/pubs/journals/resources/impact-factors). The idea behind this measure is the assumption (which is not always correct!) that journals whose articles are cited more frequently contain more important and high quality studies. Of course, this isn't always the case. As was noted in the July 27, 2016, editorial in the journal *Nature*, "*it gives disproportionate significance to a few very highly cited papers, and it falsely implies that papers with only a few citations are relatively unimportant*." Of course, as the same editorial points out, for an astute scientist, "nothing beats reading the papers and forming your own opinion" (www.nature.com/news/time-to-remodel-the-journal-impact-factor-1.20332).

On the other side of the quality spectrum we have predatory journals or "pseudo-journals." The name was originally coined by librarian Jeffrey Beall, who defined them as "*counterfeit journals to exploit the open-access model in which the author pays*," and publishers that are "*dishonest and lack transparency.*" The easiest way to recognize these journals is to see whether they require authors to pay high publishing fees, ranging from $200 to as high as $1,800. At face value, these journals may look like legitimate professional journals with a website and editorial board, claiming to offer a rigorous peer-review process. However, in reality they exist only for financial profit, publishing virtually *any* paper if the authors are willing to pay the fee. One of the reasons why more predatory journals are emerging every day is that there is a market for it: In many instances, researchers are required to publish a certain number of papers to gain tenure and promotion. Thus, some researchers could feel pressured to meet this quota and may take shortcuts, like publishing in a journal with low or no quality standards (Beall, 2012).

Therefore, if you are in doubt about the quality of a journal, check out its website (are there publishing fees?), and look for its publishing organization. In many cases, the publisher's owner is also the editor of every journal. Investigate the qualifications of the editor and the review board (if the review board is listed at all), since in many cases they do not possess the necessary academic expertise (Laine & Winker, 2017).

The Case of the "Chocolate Diet" that Duped the Public

Here is a sad example of how the predatory practices of pseudo-journals can mislead the public on a large scale. Journalist John Bohannon and several of his colleagues intentionally conducted a pseudo-scientific experiment that supposedly demonstrated the benefits of chocolate in weight loss. However, the whole point of their scheme was to show "*just how easy it is to turn bad science into the big headlines behind diet fads*." Sadly, they were right. You can find more details about the actual experiment and the events that transpired after the publication of their study in the article that John later wrote to reveal his entire scheme. (We will come back to the study when we discuss quality of academic writing.) But the moral of the story is that the scheme wouldn't have worked

if their poorly written paper wouldn't have been published in one of the predatory journals. Alarmingly, several of the 20 different so-called professional journals offering a "*rigorous peer-review process*" that he contacted agreed to publish the study, despite many glaring problems in the paper. According to John, the paper was published in less than two weeks, without reviewers' feedback, questions, or revisions, as soon as their credit card was charged. Here is the link to the summary of the case written by John Bohannon himself: https://io9.gizmodo.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800

3.5 How to Read a Scientific Paper

Let's assume you have conducted a thorough literature search, and you made sure that all the articles came from reputable journals. Now it is time to understand what they are actually telling you. Understanding academic papers is not an easy task, especially for a layperson. You may be surprised to find that even papers written on such fun psychological subject matters as "love" are often dry and technical. While it takes time and experience to develop the skills needed for reading and understanding academic papers, I will give you a few quick tips that should help to get you started reading literature on the topic of your research proposals. Of course, the ultimate goal is to be able to not only understand, but also to critically evaluate academic writing.

In the article "How to (seriously) read a scientific paper" published online in the journal Science, Elisabeth Pain lists the advice of several scientists from various disciplines who were asked this very question. I summarize their tips and include my own advice here. (Feel free to read the entire article as well: www.sciencemag.org/careers/2016/03/how-seriously-read-scientific-paper.)

- Begin with the paper's abstract. It is normally a summary of the entire article. My mentor used to say that an abstract is like a movie trailer, a quick synopsis that should help you decide whether you want to see the full feature. She was right, granted the trailer wasn't misleading you. Because an abstract typically cannot exceed 250 words, authors must choose their words wisely. Only the important elements and strengths of the paper will be mentioned and highlighted, which is what you need to quickly decide if the paper has any relevance to what you are looking for.
- If you believe the paper is relevant, or if you are still not sure, then you need to know more details. For example, if you want to know more about the topic of the study, read the introduction. It will give you a general idea of why the authors conducted their research, for example gaps, inconsistencies, or other lingering questions on the subject matter. If you are interested in the paper for its methodology, carefully examine the method section. If you want to know more about the findings, examine the discussion section. The results section is normally a lot more technical, and you will need to have a good amount of

statistical background to get a good grasp of the reported results. For this reason, if you have not taken statistics, you may want to skip the results section for now.

- I recommend creating a document where you can keep track of the articles and their findings. Briefly summarize each paper and its findings in your own words. If you find it too difficult to do, it is likely that you didn't fully understand the article. You may have to re-read the paper (minus the results).
- Finally, make a note of why you find the article to be relevant to your research project: Do the findings provide support for your hypothesis? If so, briefly elaborate on that point. Did you find the methodology to be helpful in some way to tweak the design of your study? Or did the study find something that contradicts your prediction and theory? Finally, you may want to remember the findings, as they could shed more light on the potential results of your own study.

THINK LIKE A SCIENTIST AND ASK YOURSELF THESE QUESTIONS:

Do you have enough evidence to support your predictions with relevant literature?

Was your hypothesis already tested? If so, was it supported? If so, why do you want to test it again? What is the value in conducting your study? If the hypothesis was not supported in previous studies, why?

3.6 Facts, Claims of Facts, and Opinions

Thinking like a scientist requires understanding the difference between an opinion, a fact, and a claim of fact. An opinion is a feeling, a thought, or an idea that is not based on any particular scientific evidence. It can come from your experiences or other non-scientific sources. We are allowed to have opinions, of course, but when it comes to academic work, which includes constructing facts, writing, and sharing knowledge, ideas must be based on scientific evidence.

A fact is an undisputed piece of information based on science. Here are some examples:

- The earth is round.
- We breathe air.
- People have different personalities.
- Physical illness can lead to mental illness, and vice versa.

Of course, there are a lot more facts I could list. But there are even more claims of facts—these are claims that have not gained enough scientific support to be full-blown matters of fact. The results of any one individual study are only a claim of fact, as no single study can be sufficient to elevate its findings to the level of being an undisputed fact. In essence, the difference between a fact and a claim of fact is the degree of uncertainty in the correctness of the data. At one point in history, people did not believe that the earth was round, and it was considered merely a claim of fact; but the increasing preponderance of supporting evidence eventually elevated it to the status of matter of fact.

Scientists, as you may guess, are very much sensitive (and should be) to distinguishing between opinions, claims of facts, and facts. This is evident in the way they speak about matters and in how they write. For example, their language and tone are adjusted accordingly when a statement is a matter of fact versus when it is a claim of fact. When they refer to something that is only a claim of fact, they use words such as "likely, perhaps, possibly, etc." to denote some degree of uncertainty. But when they make a factual statement, they write it as a declarative statement.

Furthermore, a fact presupposes common knowledge, and thus, doesn't need to be supported by a citation. On the other hand, a claim of fact must be supported by evidence. For example, the statement "*The earth is round*" doesn't require you to cite a scientific paper to confirm it – we all know it to be an undisputed fact. But a statement such as, "*Workplace discrimination impedes gender equality*" is a claim of fact and will need to be supported with at least one scientific paper confirming the claim.

Soon, you will have to do the same when you work on your research paper. Specifically, you will have to rely mostly on facts and claims of facts to build your case. Specifically, you will have to explain the purpose of your study by summarizing all known facts and claims of facts about the topic under investigation and by zeroing in on the unanswered questions, one of which will be addressed by your research. You will also have to support your hypothesis with some prior scientific evidence, again citing previously constructed facts and claims of facts.

3.7 Study Variables

Study Variables

The way we test a hypothesis is by observing phenomena that are being predicted. We try to see if, given the specified circumstances, our prediction will come true. In *quantitative* research, this judgment is based on the *quantity* of measurements of the observations. Our phenomena of interest could be two or several things that are hypothesized and related in some systematic fashion.

After formulating a hypothesis, a researcher should have a clear sense of what events or phenomena are to be observed and measured to test the hypothesis. And once these observations have been measured, they are now referred to as ***variables*** *because their values should vary from person to person or from time to time*. In fact, it is important that there is as much variation in the values as possible for a researcher to be able to detect predicted effects or any differences. For example, biological sex can be one of the study variables, in which case, a sample of 50 participants should vary in terms of its sex distribution—it must have roughly the same number of males and females. In summary, a variable is what gives you the data and it should vary from person to person. A typical study will work with at least two variables; and more often, a study will have multiple variables.

So let's identify the two study variables in the previously used hypothesis that *playing violent video games for an extended period of time causes* aggression. The two phenomena that will be observed, measured, and turned into study variables are *the time one plays violent video games* and *the level or the intensity of aggressive behavior*. These observations can be made concurrently (at the same time) or longitudinally (over an extended period of time). And, if the results reveal that the increase in time spent playing violent video games is associated with an increase in aggression, researchers can conclude that the hypothesis has been supported.

Dependent (Criterion) and Independent (Predictor) Variables

Recall that one of the key philosophical principles of science is determinism, which is the idea that any predictable event has one or several underlying causes. So, when a hypothesis is made about a relationship between two or more events or behaviors, the assumption is that one of the events may be the cause of the other. The variable that is hypothesized to be the cause is called the ***independent variable*** because its value does not depend on the other variable(s) of the study. The variable that is predicted to be caused by the independent variable is referred to as the ***dependent variable*** because its value *does* depend on the behavior of/change in the independent variable. A study that seeks to test such a cause-and-effect relationship, known as an ***experiment****,* will thus always have a dependent and an independent variable. Studies that are not designed to demonstrate cause and effect, referred to as ***correlational****,* will still usually have at least two variables; but only an associational relationship between them can be investigated. Although correlational studies cannot test causal relationship between variables, they can still hypothesize about whether or not one of the variables contributes to or consistently precedes the occurrence of the other. The variable that predicts or precedes the other variable is called a ***predictor***, and the variable that is predicted is called a ***criterion*** variable.

Extraneous and Confounding Variables

Even when a hypothesis has a clear dependent and an independent variable, researchers do understood that human behavior and mental processes are very complex and are usually affected

by multiple factors. In fact, the potential multitude of biological and social factors influencing any given behavior makes psychology a challenging but also an exciting enterprise. By systematically observing a phenomenon, a researcher is able to focus on the effect of variables of interest and control other factors with various techniques. Any factors that are *not* specified by the hypothesis of the study but can still potentially have an effect on the dependent variable are referred to as **extraneous variables**. Reviewing the literature can help identify them; and if possible, they can also be included in the study and controlled. This scenario will be discussed next.

Confounding Variables. Extraneous variables that are known to exert some effect on the dependent and independent variables are called ***confounding variables***. They are termed confounding because, if not included in the study, they will most likely confound or distort the results of the study and undermine its validity.

To illustrate this let's go back to the article by Anderson and Dill (2000), where the focus was on testing a potentially causal link between playing violent video games and aggression. In the introduction of the article, the authors made it clear that the hypothesis about the causal link between playing violent video games and aggression was partially fueled by the 1999 Columbine High School shooting tragedy, when two high-school students, who reportedly liked to play violent video games, killed and wounded several students from their school for no apparent reason. The media and people alike couldn't help but wonder about the link between the two facts : the teens committed violent acts, and they reportedly liked to play violent video games. So, it seemed logical to draw a connection between the two variables and to hypothesize that the frequent exposure to violence through video games could have caused or contributed to the teens' extreme violence (see Figure 2).

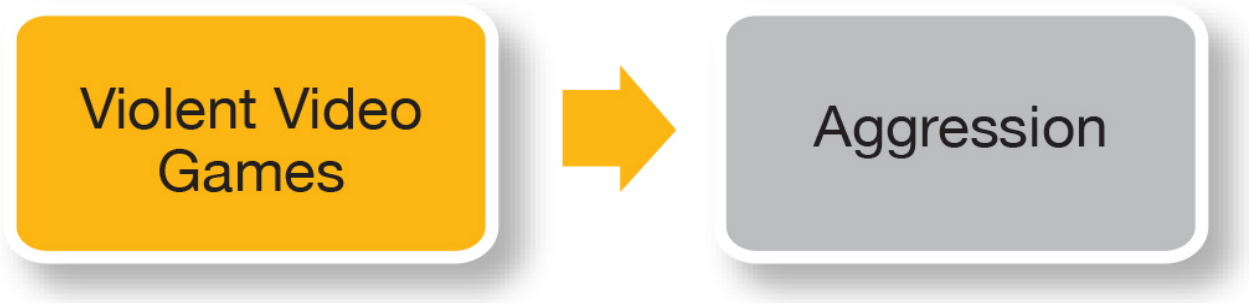


Figure 2. *Explanation #1: Violent Video Games Cause Aggressive Behavior.*

However, the problem with simply testing this hypothesis is that there are several ***extraneous*** variables that can also have an effect on the dependent variable (i.e., aggression); and if they are not considered, the results of the study can be confounded by their potential influence on the outcome. For example, having an antisocial personality disorder or a

violent personality trait can drive a person to act violently and also to be more inclined to play violent video games. This seems like a very plausible *alternative* explanation #2 (see Figure 3).



Figure 4. *Alternative Explanation #3: A person with a violent personality trait is more influenced towards violence by playing violent video games.*

Notice, all three possible explanations can be inferred from the previous studies on video games and aggression. Thus, personality is clearly a confounding variable and, at a minimum, should be included in a research study to account for its potential influence on aggression. In Anderson and Dill's (2000) study, personality was included and measured by giving the participants a survey; the authors hypothesized that aggressive individuals would be more strongly affected by playing video games by having more aggressive thoughts and a higher affective arousal than the typical players.

Known and unknown extraneous variables can be also controlled experimentally, as you will later see, by randomly assigning participants to control and experimental conditions; this strategy is both effective and efficient--It distributes any potential differences in participants' personality or background evenly (because it is done randomly) between the groups and it doesn't require measuring all those potential differences (As they are evenly distributed across the groups; and will affect the dependent variable equally in all the groups).

In the example with violent video game and aggression, the confounding variable personality is likely to distort the results by demonstrating a false connection between the independent variable video games and the dependent variable aggression. But that's not the only way that a confounding variable can distort the results. For example, a team of researchers, led by Drs. Gunter and Murphy, conducted a large multinational study to examine a potential association between drinking coffee and mortality. During interviews or in surveys over five hundred thousand participants reported the number of cups they consumed per month, week, or day, along with information on lifestyle, such as, education, smoking, alcohol consumption, physical activity. Researchers also obtained data on participants' health and mortality. As it turned out, smoking was associated with both the independent variable drinking coffee and mortality. Specifically, people who smoked were more likely to drink more coffee and also more likely to die from several diseases. At the same time, results also showed that drinking more coffee was associated with lower risk for death (some cancer types and cardiovascular disease for women). After controlling (statistically) for the effect of smoking, the positive association between drinking more coffee and lower risk of death was even more increased, confirming the predicted association and apparent health benefits of coffee. Gunter et al. (2017).

Constructs and Operational Definitions

As was mentioned previously, all variables in a quantitative research are measured using numbers. This is easier to accomplish when the behaviors under investigation have tangible physical characteristics and can be detected and measured using objective instruments. A good example of this is our sensory perception—its neural processes can be detected and measured by a number of different instruments, for instance, by Positron Emission Tomography (PET), functional Magnetic Resonance Imaging (fMRI) and/or Event-Related Potentials (ERP). However, more often than not, researchers are interested in studying phenomena that are *non-tangible* and *abstract.* These are known as ***constructs***. Some of the commonly studied constructs in psychology are love, intelligence, happiness, and attitudes. And, of course, there are many others.

The way that psychologist can circumvent the vagueness and intangibility of an abstract phenomenon and still be able to study it scientifically is by developing an ***operational definition****,* a definition that describes a construct in quantifiable and observable terms or properties. In addition to an operational definition, a researcher can use a ***coding scheme***, which is a more elaborate version of an operational definition, describing all different possible variations of the characteristics that a given construct can take on and their numerical values. For example, an operational definition of sensitive parenting can be the following observable behavior: (1) timely response to a child's bid for attention, (2) expression of positive emotions in response to a child's expression of the same, and (3) response of empathy to a child's expression of sadness. Using a theory or prior empirical evidence as guide, a researcher may further specify that all three behaviors have to occur in a given time to constitute sensitive parenting, or the same researcher

may operationalize it by establishing the following rule: the occurrence of one of the three behaviors is sufficient to consider it to be sensitive parenting.

3.8 Conceptual vs. Operational (Statistical) Hypothesis

A distinction should be made between an operational and a conceptual hypothesis. From the previous section, you have learned that researchers oftentimes deal with constructs that do not possess physical properties. Thus, to study such constructs researchers use operational terms in order to define and measure them. When a hypothesis is stated using such operational terms, it is said to be an ***operational hypothesis*** (also known as a ***statistical hypothesis***). On the other hand, a ***conceptual hypothesis*** is stated in abstract terms and is sometimes referred to as an "educated guess," predicting or conceptualizing the relationship between a set of variables. Somewhat similar to the relationship between a hypothesis and a research question, an operational hypothesis guides a study by specifying the terms and the context within which the variables are going to be measured. While a conceptual hypothesis helps the researcher see the study findings within broader context.

For example, a researcher may conceptually hypothesize that playing video games can have cognitive benefits. Next, to test her conceptual hypothesis, the researcher will have to operationally define the construct of interest in order to test it. In this example, the construct is ***cognition or cognitive benefits***. She can operationally define it in several ways, for example by focusing on attention allocation, spatial resolution, or mental rotation ability. Thus, the researcher can test one or all three of these operational hypotheses:

1. Video game players will have more accurate and/or faster attention allocation.
2. Video game players will demonstrate a higher spatial resolution in visual processing.
3. Video game players will have increased mental rotation abilities.

Then, in her discussion, she can come back to the conceptual hypothesis to explain how the findings of her study fit within the broader question about the link between video games and cognition.

THINK LIKE A SCIENTIST:

As you think about your research project, begin by identifying your research goals and stating your conceptual hypothesis. Then, think of ways you can operationally define the terms in your conceptual hypothesis and test an operational hypothesis.

3.9 Measures of Scale

Measures of Scale

It is important to know the measurement scale of the variables for at least two reasons. First, it will determine the type of the statistical analysis that one can conduct to evaluate the data (we will come back to this point in the chapter on descriptive and inferential statistics). Second, if a researcher uses a coding scheme to measure variables, the coding scheme will impose a certain measurement scale, which will, in turn, affect your statistical analysis. In the following section we will review four commonly known scales of measures—nominal, ordinal, interval and ratio—and discuss their properties.

Nominal Scale

The sole property of this scale is that it categorizes things or people into two or more categories based on particular characteristics. For example, gender is a nominal variable, because it categorizes people into states of male or female. Political affiliation is another example of a nominal variable because it classifies people by their membership in political parties. It's important to point out what this scale does not do. It doesn't differentiate people or things of different categories quantitatively. For example, someone who belongs to a Republican political affiliation category does not differ quantitatively from his/her Democratic counterpart. The only difference between people belonging to these different categories would be in the ***kind*** of political affiliation they belong to.

Because this scale doesn't assign meaningful quantitative values to observations, researchers are limited in the kinds of descriptive and inferential statistical analyses they can accomplish with nominal variables. And while some variables are inherently nominal, e.g., gender, so the scale can't be changed, other variables can be measured using a more meaningful measurement scale. Nominal scale can have multiple categories (more than two categories). For example, categorizing people into five different personalities is a nominal scale with five different categories.

A subcategory of a nominal scale is a **dichotomous** scale, which means a scale with only *two* points that are polar opposites of each other. For example, gender can also be called a dichotomous scale if it includes female and male categories only. Another example is an absence or a presence of some thing—these are two polar opposites states.

Ordinal Scale

Similar to a nominal scale, a variable measured on an ordinal scale places items or people into categories. Additionally, it rank-orders them; for example, from high to low or from big to small. Public colleges and universities could be categorized ***and*** rank-ordered by the quality of their education, from 1 to 100 (with 1 being the top-quality school), as based on a chosen ranking scale.

However, this scale doesn't include the difference in the magnitude between ranks; in other words, there is no numerically meaningful difference between the ranks in a scale. For example, let's use the *U.S. News and World Report* 2015 best public colleges' ranking (see http://colleges.usnews.rankingsandreviews.com/best-colleges/rankings/national-universities/top-public; 2015, July 24) to demonstrate this point.

The ranking of public colleges used by *U.S. News and World Report* ranges from 1—college of the highest quality—to 117—the lowest category on the scale. The number one spot in this ranking is given to University of California at Berkeley (#1), followed by University of California, Los Angeles (#2), and University of Virginia (#3). But how different *is* the University of Virginia from the top ranked University of California at Berkley? Similarly, is the difference between first and second place in the rankings the same as the difference between 3rd and 4th or 5th and 6th places? The rankings do, in fact, contain that information. So, while the ordinal scale provides more information than the nominal scale, it is still limited in its properties, and thus, is limited to the kinds of statistical operations that can be done with these types of variables.

Many psychological rating scales are also ordinal. For example, rating something on a scale of 'bad,' 'good,' and 'great' has an implicit order, but it doesn't tell us the numerical distance between 'bad' and 'good' and between 'good' and 'great'. The very popular 7-point Likert scale is also an ordinal scale, despite the fact that its response anchors are equally distributed as these are simply arbitrarily created points using real numbers.

Interval Scale

A variable that is measured on an interval scale contains information about the order and also the distance between the ranks. Thus, for example, if the distance between the ranks of a scale is numerically equal and meaningful, it is an interval scale. A good example is the measure of temperature; whether it is measured in Fahrenheit or Celsius, the scale slides from freezing to boiling points in equal intervals or degrees. A 10-degree difference in the temperature from 40 to 50 and from 60 to 70 has the same physical property. When the temperature drops from 33 degrees to 32 degrees Fahrenheit, this numerical change in the temperature carries a meaningful physical change—water reaches the freezing point.

The limitation of an interval scale is that it does not have a true zero. For example, zero degrees Fahrenheit doesn't mean an absence of temperature. However, it can still be arbitrarily made meaningful. For example, zero degrees Celsius is meaningful in that it represents the freezing point of water; however, this meaning of zero is arbitrary because we know that the freezing point in Fahrenheit is 32 degrees. Although an interval scale is not as sophisticated as a ratio scale, it can be used in the same statistical analyses as the variables measured on a ratio scale.

Ratio Scale

A ratio scale shares the same properties as an interval scale, as well as the three other scales, but it also has a true zero that represents a true absence of whatever is being measured. A couple of examples of ratio scale variables are years of age and years attending university. 'Zero' may represent the absence of age, or number of years of college education. In either case, 'zero' corresponds to the absence of any accumulation of years.

Important Concluding Remarks: As you are planning your study, make sure you have a testable hypothesis that is linked to a broader research question, and that addressing this research question would contribute to a larger body of knowledge. Your hypothesis has to be grounded in theory and/or current empirical evidence. You have to identify your study variables and their measures; the constructs have to be operationally defined. This process—from formulating a hypothesis to operationally defining the study variables—has to be informed by prior studies and theories, so you need to conduct a thorough literature review to justify every decision you make in the process. These decisions, including the scale of the chosen measures, will affect the subsequent stages of the research process—***data collection***, ***statistical analyses*** **and** ***interpretation of results***.

3.10 Brief Overview of Different Research Designs

Goals of Science

I began this book by trying to convince you that science offers the most reliable pathway to new knowledge and to explaining unknowns. Acquiring knowledge is, of course, the main reason for scientific explorations and the goals of these explorations can be broken down into the following three broadly defined categories: ***describing***, ***explaining***, and ***predicting***. Describing a phenomenon is the first step towards unraveling its mystery. For example, before developmental psychologists can understand what causes language delays in some children, they have to know what the normal and age-appropriate language skills are and their normal rate of development; i.e., someone has to *describe* the normal stages of language development. The second goal is to explain the phenomenon of interest. This is a more challenging task than simply describing a phenomenon because it requires knowledge of its causes. And, because understanding causes contributes to more accurate explanations and predictions and better control, most scientists consider discovering causes to be the ultimate goal of science. Still, we must recognize that all three goals are interconnected and important. Being able to distinguish normal from atypical development, for example, can help researchers locate the causes of developmental problems, which, in turn, can lead to early intervention and prevention.

Types of Research Designs

Recall that what distinguishes science from other forms of inquiry is the reliance on systematic observations. By systematic we mean a certain set of rules and principles that will yield the most accurate and unbiased results possible, usually by directly observing the effect of some hypothesized cause on the behavior of interest. But, given that the primary subjects of interest in psychology are usually humans, researchers may be restricted in how these observations can be accomplished. For example, a researcher is unable to deprive children of a loving home just to test the effect of deprivation of love on child well-being—researchers have to use other methods. Thus, multiple research methods are available to allow researchers find ways to observe phenomena of interests and collect data.

The term ***data*** refers to observations that have been measured and become study variables. While there are many research designs, only one method of data collection can address questions of causality: an ***experimental design***. And since finding causes is, arguably, the ultimate goal of science, all research designs are grouped into ***experimental***, ***quasi-experimental***, and ***non-experimental***, based on their ability to address causal questions (see examples of different research designs in Table 1). The next two chapters will cover two non-experimental designs: content analysis and observational. Two experimental designs—within- and between-subjects—will be covered next. Finally, after you grasp the main principles of the experimental methods, we will discuss quasi-experimental designs—developmental/longitudinal and nonequivalent group (see Table 1). Notice, these methods we will cover are not, by any means, an exhaustive list of the variety of methods psychologists and behavioral scientist utilize in their research. You can find a few other, some even quite familiar to you, examples listed in table 1.

Table 1. Examples of Different Research Designs

Research Designs	Examples
Experimental	Single-case
	Within-subjects
	Between-subjects
	Mixed designs
Quasi-Experimental	Developmental/longitudinal
	Nonequivalent group
	One-group design
Non-Experimental	Observational
	Content Analysis
	Survey

Establishing Internal Validity Through Control of Variables

The degree to which a study can ascertain cause and effect refers to its ***internal validity***. The best way to demonstrate cause is by controlling its variables—(1) systematically manipulating the independent variable and (2) controlling the effect of other extraneous variables by either holding them constant or eliminating their effect. In other words, if you believe that some variable is causing an effect and you need to test it, manipulate the cause to observe its effect. However, to be absolutely sure that the effect came only as a direct result of your manipulation, and not from other extraneous and confounding variables, all other potential causes have to be kept unchanged. An experiment is a type of research design that can both establish the maximum level of control and test cause and effect, and therefore, has the highest internal validity.

Establishing External Validity

While demonstrating cause is the ultimate goal for many, if not most, researchers, it comes at a price—some loss of ***external validity***, which refers to the degree results of a study can be generalized to real life. Artificialities, created by highly controlled experimental conditions, make it difficult sometimes to apply the results to real people and/or real life circumstances, and thus, may undermine the whole effect of a study. So, there is a constant tradeoff between internal and external validity. An experimental research design delivers the highest internal validity, but has

the lowest external validity. On the other hand, non-experimental design has the lowest internal validity but can achieve the highest level of external validity. Quasi-experimental designs can be placed somewhere in between.

Sample vs. Population of Interest

The population of interest is about *what* or about *whom* the results of research are being generalized. In psychology, this is normally, but by no means exclusively, people at large. Because psychology is the study of the *human* mind and behavior, many research questions concern the human populace. To give an example, researchers Farrant and Zubrick (2011) studied the importance of parent-child book reading and joint attention (visually sharing or following partner's attention) during infancy on the development of children's vocabulary three years later. The population of interest in this case is all normally developing children and their parents because the results will pertain to all children and their parents if they communicate, read books, and share joint attention. Other times, the population of interest may be limited to a certain group of people or animals. For example, Hoff, Core, Place, Rumiche, Senor, and Parra (2012) compared the rate at which bilingual and monolingual children learn their native language(s) in order to better understand the typical development of bilingualism. The population of interest, in this example, was the bilingual children and their parents.

Thus, a representative sample increases the external validity of a study.

Random Sample Selection

Since it is important that the selected sample is representative of the entire population of interest, it must be selected randomly. Random sample selection means that every individual, animal, or entity (if it's a content analysis) has an equal chance of being selected for the study. Listed below in Table 2 are some of the most commonly applied sampling techniques.

Table 2. Sampling Techniques

Types of Probability Sampling Techniques	Description	When to Use
	Probability of selecting a participant/subject/entity is known. This form of sampling yields a random sample.	
Simple random sampling	When the probability of being selected is known. The process starts by listing all members or units of the target population. Next, using a random procedure method (e.g., randomly selecting people from a list) select the sample. You can use free random number generators on the Internet. For example, https://www.randomizer.org/. This strategy is not common.	Good for a small population.
Systematic sampling	Similar to simple random sampling. The process starts by listing all members or units of the target population. Next, select every kth person/unit from the population list. "K" is the interval between the persons/units. For example, every 10th or every 100th person on the list. The decision for the k-selection must be justified.	Good for a small population
Cluster sampling	Identifying clusters of participants/subjects/entities first; then selecting only a few of the clusters (randomly or most representative of the population) and including all of the participants/subjects or entities from those clusters.	Use it when some clusters cannot be included, or there are too many clusters to include.
Stratified random sampling	Dividing the population into groups (i.e., “strata”) and randomly selected participants from each strata.	Use it when it is particularly important to have a representative sample of different groups (e.g., minority groups)
Convenience Sampling	Probability of selecting a participant/subject/entity is not known. Examples of this are using volunteers or people who are readily available to a researcher (e.g., college students). Use it when non-random sampling can be theoretically or methodologically justified.	

References

Anderson, C. A. Ihori, N., Bushman, B. J., Rothstein, H. R., Shibuya, A., Swing, E. L., Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in Eastern and Western countries: A meta-analytic review. *Psychological Bulletin, 136*, 151–173.

Bretherton, I. (1992). The origins of attachment theory: John Bowlby and Mary Ainsworth. *Developmental Psychology, 28*, 759–775.

Farrant, B. M., & Zubrick, S. R. (2011). Early vocabulary development: The importance of joint attention and parent-child book reading. *First Language*, 1–22.

Gunter, M. J., Murphy, N., Cross, A., Dossus, L., Dartois, L., Fragherazzi, G., …Riboli, E. 2017. "Coffee drinking and mortality in 10 European countries: A multinational cohort study." *Annals of Internal Medicine*: 1-12.

Hoff, E., Core, C., Place, S., Rumiche, R., Senor, M., & Parra, M. (2012). Dual language exposure and early bilingual development. *Journal of Child Language, 39*, 1–27. DOI: 10.1177/0142723711422626

Moll, H., & Meltzoff, A. N. (2011). How does it look? Level 2 perspective-taking at 36 months of age. *Child Development, 82*, 661–673.

Sanbonmatsu, D. M., Posavac, S. S., Behrends, A. A., & Moore, S. M., & Uchino, B. N. (2015). Why a confirmation strategy dominates Psychological science. Plos One, 10.

Svanberg, P. O., Mennet, L., & Spieker, S. (2010). Promoting a secure attachment: A primary prevention practice model. *Clinical Child Psychology, 15*, 363–378.

Chapter 4: Content Analysis

Introduction



Content analysis is a research technique that is used to systematically analyze content of any meaningful body of communication (e.g., pictorial, verbal, written, etc.) with the purpose of understanding its symbolic meaning, making inferences and predictions. It is also one of the oldest methods that has been utilized not only in psychology but in a variety of other academic disciplines, such as anthropology, history, public administration, and the political sciences, as well as in nonacademic settings, such as program evaluation. In this chapter, I will focus only on the process of content analysis as it applies to ***quantitative*** research. Unlike qualitative research, where extracting and describing categories or patterns is the end result, quantitative research applies the analytic procedures of content analysis to further transform the extracted categories into numerical data, which then is used to test hypotheses with statistics.

4.1 Content Analysis in the Process

The following is a simplified version of the process that goes into content analysis:

- Formulating a research question and specifying a hypothesis
- Providing theoretical, empirical, or logical rationale for selecting specific qualitative data and for conducting content analysis
- Selecting a random sample (determining what will be the unit of sample)
- Transforming constructs into study variables
    - Developing conceptual and operational definitions for study constructs
    - Selecting a unit of analysis
    - Choosing between time and event sampling process of coding the data
        - Selecting a unit of coding (for time sampling) and a unit of context (optional)
    - Assigning codes/numeric values to qualitative categories (e.g., "1," "2," etc.)
- Transferring coded data/numeric values into a statistical program for statistical analysis

To illustrate this process we will review the following studies:[1]

- Study 1: GAO Report (2012), Returned Peace Corps Volunteers
- Study 2: McAdams, Diamond, Mansfield, and Aubin (1997), Stories of Commitment: The Psychosocial Construction of Generative Lives
- Study 3: Mastro and Stern (2003), Representations of Race in Television Commercials: A Content Analysis of Prime-Time Advertising
- Study 4: Klimenko (2011), Mother-Toddler Intersubjectivity as a Contributor to Emotion Understanding

Formulating Research Questions and Specifying Hypotheses

Any study begins with formulating a research question and choosing an appropriate research design that can address it. Content analysis is no exception; it needs to be appropriate for the questions a researcher sets out to address.

[1] *The full citations of these studies are provided in the reference page at the end of the chapter.*

Sources of Qualitative Data



Qualitative data is all around us; it is nonnumerical information that can come from a variety of sources. This data may originate from a carefully designed study that yields qualitative information, from a study that was designed by another researcher and for a different research purpose, or from public or private records of government or private institutions. For example, in a study by McAdams, Diamond, Mansfield, and Aubin (1997; study 1), researchers compared the life story narratives of 70 older adults to find out whether the stories of high generative adults differ from the less generative adults, with ***generativity*** defined as having concern for the well-being of the next generation. The adults' narratives, in this example, are the qualitative data that will have to be transformed into numbers to make it suitable for statistical analyses. ***Study 4*** is an observational study where 79 mothers and toddlers were observed interacting during a book-reading session.[2] The purpose of this study was to measure the degree of ***intersubjectivity,*** or the shared understanding between two or more people, between mothers and their young children in order to test a hypothesis that the extent to which mothers and children shared intersubjectivity would predict children's emotion understanding two years later. Intersubjectivity, marked by

[2] *Results of this study were used for my dissertation.*

mutual gazing and affect sharing, is also an example of qualitative data that was derived from a carefully designed observational study (i.e., Beebe & Lachmann, 1988).

Qualitative (and quantitative) data that was collected by someone else or has naturally accumulated is referred to as ***archival data***. For example, the US Census Bureau is a government institution that collects quantitative data about the US economy and its people. These data are open to the public and can be used for academic purposes. A researcher who chooses to use this type of data is said to be using archival, quantitative data collected by the government. Television advertisements can also be considered qualitative data, because they carry meaningful messages. These messages visually and verbally manifest our cultural and societal norms and beliefs and, as such, can be systematically evaluated by a researcher to learn more about our society and its social issues. An example of such research is a study by Mastro and Stern (2003), who have systematically evaluated portrayals of characters of different racial and ethnic backgrounds in television advertisements to test the hypothesis that minority characters are portrayed in different contexts and with different frequencies than white characters (***study 3***). The authors argued that these differences are likely to have an effect on the self-perception of the viewers who are likely to identify more closely with the depicted characters of the same racial and ethnic background.

Theoretical, Empirical, or Logical Rationale for Content Analysis

In the Absence of Quantitative Data.

Content analysis can be time consuming and limited in the kinds of inferences it allows researchers to make. So, why would someone choose this method of analysis? Sometimes, the rationale to use qualitative data for content analysis could be as simple as the fact that qualitative data may be the only available data an analyst has at his/her disposal. Historians, for example, use past documents and narratives to piece together historical events just because the nature of their inquiries limits them to using information that is mostly left by our past and generated in unsystematic fashion by people in the course of their lives. Program analysts, whose job is to evaluate the merits of government or private programs, sometimes have to resort to the analysis of qualitative data (e.g., organizational documents, staff interviews) in order to determine the quality of the program. A good example is the US Government Accountability Office (GAO), a federal agency whose primary function is to conduct evaluations of the federal programs at the request of the US Congress. One of the agency's reports, which we will review, was based on the content analysis of the denial of health care coverage letters—an example of qualitative data. This analysis was then used to ascertain the main reasons (categories) for the denial of benefits.

When the Content *is* the Focus of the Study.

Other times, the choice to use content analysis is driven purely by the desire to better understand the content itself: when it can reveal people's attitudes, mental states, or intentions, etc. For example, in my study on mother-child intersubjectivity, I predicted that the frequencies with which mothers' and children's facial expressions and eye contact were matched (study 4) would reveal their emotional connection and quality of their relationship.

Another example is when researchers need to understand the content of the phenomenon in question in order to theorize about how this phenomenon can affect thinking and behavior. For instance, researchers need to know the content of video games before they can make any assumptions about how video games can affect gamers' state of mind.

An example of a more practical application of content analysis is a study comparing online course discussions with similar but in-class (face-to-face) discussions to see which forum generates more higher-order (critical) thinking (e.g., Marra, Moore, & Klimczak, 2004). In this case, the focus of the study is specifically on the content of the generated discussions in two different forums.

As Converging Evidence.

Although this occurs less frequently than with other research designs, content analysis can sometimes provide additional supporting evidence for ideas that have been supported by other studies but lack some external validity. Recall that no one particular type of research design is perfect—a researcher always has to balance between external and internal validity. An experiment is the only method that can test causal theories, and thus, this design tips the scale toward higher internal validity at the expense of lowering the external validity, or its ability to generalize the findings to the general population. In correlational research, which includes a content analysis, the scale tips in the opposite direction, against internal validity and toward external validity. Since content analysis is especially useful for evaluating naturalistic and sometimes unstructured real-life data, it can provide more evidence, and with a higher degree of external validity, for something that has been found in artificial settings or needs more real-life supporting evidence. For example, in a study by Gerhards (2010), a survey of European citizens (the 2000 European Values Survey) showed that people in Turkey were less likely to accept "homosexuality as justified" compared to other European countries. Although this finding suggests that Turkish citizens hold more negative attitudes toward gays and lesbians, more evidence is still warranted.

One way to find more supporting evidence, to expand our understanding about Turkish culture, and, perhaps, to find the roots of such negative attitudes would be to examine Turkish TV programming and compare the depiction of homosexuality in Turkish TV shows with those of other countries with more favorable attitudes toward gay and lesbian people. Given the findings in Gerhards (2010), one can hypothesize that Turkish TV programs may completely lack gay or

lesbian characters in TV programs or portray them in a negatively stereotypical fashion. If this hypothesis is supported, it will provide more converging evidence for the conclusions of Gerhards (2010).

A study conducted by Tse, Belk, and Zhou (1989) is another great example of how content analysis can be a useful design in providing converging evidence. Specifically, the authors of the article systematically analyzed the newspapers' consumer ads printed between 1979 and 1985 in three Chinese societies: Hong Kong, the People's Republic of China (PRC), and Taiwan. The purpose of this cross-cultural and longitudinal examination was to better understand the evolution of consumer societies, and, specifically, to test a theory that differences in consumption behavior are partially driven by the consumption values of these societies. Thus, for example, Tse et al. (1989) hypothesized that the PRC, which was largely a utilitarian-based society, would show more consumer appeal toward product performance, quality assurance, and technological content ad themes, presumably because of the people's higher risk aversion due to such cultural values as collectivism and self-sacrifice. Whereas, the people of Hong Kong, identified as a hedonistic society by the researchers (e.g., at the time the article was written, the people of Hong Kong had higher income and were more experienced consumers as compared to the other two countries), would most strongly resemble the behavior of the US consumers and, thus, would show more appeal toward luxury goods. The authors pointed out that the literature on consumer revolution had mostly been examined from a historical point of view and that their content analysis would be the first to provide an important contemporary perspective on this issue.

Another interesting example of how a content analysis can complement other research is the study of two cities, Boston and San Fransisco by Plaut, Markus, Treadway, and Fu (2012). The purpose of this content analysis was to test a broader theory that the ecology and the history of a city influences people's values and norms residing in those cities. They focused on two such cities, Boston and San Fransisco. While the two cities are similar in many ways, e.g., they both are highly ranked in terms of hi-tech industries, both have a high concentration of universities, and have high median income, their different settlement histories make them also different in their adherence to norms and traditions. Specifically, Boston is the "old and established" city with longer history, whereas San Fransisco is the "new and free" area. Thus, the longer history of Boston will produce the norms of "following the norm" and old traditions; and the latter will produce the norms that value uniqueness and independence; and adhere less to old norms and traditions. The authors first surveyed the residents of both cities, measuring their perception of their respective city's norms. For example, one of the survey questions asked to rate their agreement to the following statement: "In this area, there are very clear expectations for how people should act in most situations". As the researchers predicted, people of Boston perceived more clear norms in their area than the people of San Fransisco. To compliment the survey results and to provide more support to their original theory, they content analyzed and compared themes in newspaper headlines of Boston Globe and San Francisco Chronicle, and the corporate statements of venture capital firms and

hospitals of the two cities. These analyses showed that novelty and freedom was more predominant in San Fransisco, while establishment and tradition was more prevalent in Boston.

In summary, whatever the purpose of content analysis may be, a researcher will have to conduct a thorough literature review on the topic of his/her interest and then demonstrate the suitability of content analysis to address the goals of the study.

Manifest vs Latent Content

Typically, images and texts can be analyzed in terms of their literate content (i.e., manifest) and/or hidden (i.e., latent) content. Take, of example, a photograph that a friend of yours might have posted on Facebook or Instagram. You may evaluate the photograph based on who and what is depicted on the photo, e.g., number of people, clothing, location, etc. By focusing on the literal details of the image, you are essentially engaged in a manifest analysis. But in addition to that, most all of us, evaluate a photograph based on its latent content—e.g., its hidden messages. Similarly, when conducting a content analysis, an investigator may focus on a manifest or a latent content, or a combination of the two. Manifest content can be viewed as more objective as it is based on observable features, such as a number of people, explicit words used in a text, etc. While latent content is typically more subjective as it implies "reading between the lines" (Holsti, 1969) so to speak. The same image maybe interpreted differently by different people. Let's go back to the example of a friend of yours who posted an image of herself/himself in some exotic location. You may interpret it as bragging, while someone else may see it as an expression of enjoyment and happiness. Despite it being potentially subjective, latent analysis can be quite revealing as the goal of latent analysis is to dig deeper into the "true" meaning or nature of the content. In other words, we can conceptualize the manifest as the surface and the latent as the deep structure of the content (Bengtsson, 2016).

Let me give you an example of a study that looked at manifest and latent content of photographs of Russia taken by American and Korean tourists (study by Kim & Stepchenkova, 2015). The point of the study was to see if the latent content would be more revealing in terms of predicting people's attitudes and willingness to travel to Russia after seeing the images.

The manifest content were categories of people (e.g., single or groups), nature landscape (e.g., urban or rural), place (e.g., tourist or residential), etc. As you can see, these were all straightforward details found in the images. Some of the examples of the latent content were friendliness (vs unfriendliness), uniqueness, and pleasantness (vs unpleasantness). The results revealed that both manifest and latent content influenced people's willingness to travel to Russia. Among the latent features that affected the desire to visit Russia were cleanliness, friendliness, and uniqueness.

In most all psychological research, manifest and latent content are interrelated: That is, deeper structures of our inner-psyche (e.g., personality, feelings, intentions, etc.) are manifested in some in observable way on the surface. For example, according to discrete emotions theory, people's explicit facial expressions, such as genuine smile, reflects feelings such as happiness or enjoyment. And so, researchers can use facial expressions (i.e., manifest content) to make inferences our people's feelings (i.e., latent content). Notice, that such assumptions should be based on some theory or prior evidence. The studies that will be described below will use manifest content to make broader assumptions about the latent or hidden information.

Selecting a Random Sample, Choosing the Coding Process, and Transforming Constructs into Variables For Statistical Analysis



Once you have established your research goals and you have clear hypotheses in mind, the next step is to develop a plan for how the phenomenon of your interest should be observed and measured. A researcher will have to decide “who” or “what” will be able to provide him/her with the information or the measure of the phenomenon, and this group of randomly selected people, subjects, or entities will be the sample—the unit of sample. In a typical psychological study, people are the population of interest, and they are selected at random to provide information about the phenomenon a researcher is interested in measuring. In content analysis of quantitative research,

however, a sample is usually not a conventional sample of participants; instead, a sample could be a collection of documents, interviews, TV commercials, pictures, or any other artifacts that can tell the researcher something about the people and the topic of interest.

In content analysis the phenomena of interest are typically concepts or abstractions that are intangible, have no physical properties, and cannot be directly observed. These are known as *constructs,* and will typically represent the *latent* content. But since the quantitative approach involves testing hypotheses with numbers and statistics, researchers have to find ways to measure these constructs in mathematical terms.

Choosing Between Event and Time Sampling Process of Coding the Data

Broadly speaking, there are, at least two ways, that behavior or phenomenon you are examining and coding, can be "coded". First, you can choose to use *event sampling* approach. This is when you record all instances of the behavior/phenomenon of interest within the entire time frame of your study. For example, if you are examining instances of depicted moral themes related to group loyalty within an entire film, using event sampling process entails recording every depiction that you observe during the viewing of the film. The second approach is called *time sampling*, and it entails choosing a very short interval of time (a unit of coding) during which you will record if the behavior/phenomenon of interest has occurred or not. Thus, if you coding moral themes of loyalty in a film, you could choose to use a 5-minute of unit of coding, and you will record (pause and record) every 5 minute intervals the presence or the absence of the theme. The unit of coding is arbitrary but should be thoughtfully chosen based on the frequency and the duration of the expected behavior/phenomenon. This time interval should be short enough so that, ideally, the behavior/even you are coding does not happen more than once within it; but it should also be long enough to make sure that there is enough time for the behavior/event to run its course.

Put Everything Together Into a Coding Scheme

The following steps (a–d) represent a process of developing a *coding scheme* which is typically based on using observable (i.e., manifest) content to obtain measurable data of qualitative constructs and latent content. It is worth noting that this process doesn't always follow the steps in the order from *a* to *d*. For example, in some cases, it is easier to determine the unit of analysis first and then move on to establishing the coding process and operationally defining constructs. In other cases, a researcher knows the constructs he/she would like to observe but may need more time to determine what coding approach will best fit with the study (event vs time sampling) . What helps me in this process is creating a data table with the columns and rows where the unit of analysis and variables should go.

a. Provide a conceptual and an operational definition of the study construct(s). The *conceptual* definition is expressed in theoretical/abstract terms, while the *operational* definition is expressed in concrete and measurable terms—in terms that can be directly observed and measured. The operational definition has to match the conceptual definition as closely as possible. In some cases, an operational definition is no more than a list of unambiguous and observable categories (as you shall see in the study 1 example).

b. Next, identify the *unit of your analysis*. This is what you want to analyze and from which you will draw conclusions. But a more accurate way to describe it is to say that it is who or what will provide you with the data. Usually, but not always, it is the same as your unit of sample. In a typical psychological study for example, each tested individual is the unit of analysis and also the unit of sample. Thus, for example, a psychologist who studies personality will collect a sample of people who will be the unit of sample and the unit of analysis, since each participant will have some sort of data on their individual personality. In content analysis of TV commercials, for example, each commercial can represent a unit of sample and a unit of analysis if an entire commercial provides a researcher with the measure of the phenomenon of interest (e.g., emotional valence of commercials). Alternatively, suppose a researcher is interested in characters' emotional portrayal. Then the unit of analysis is going to be the characters in the sampled TV commercials.

c. Select the coding approach: event or time sampling. If you choose to proceed with time sampling approach, determine your unit of coding.

d. *the unit of coding* and*the unit of context*. These two terms set the boundaries of the portion of the unit of analysis that is to be coded to measure the construct of interest. Typically, they are needed when a unit of analysis or the phenomenon that it is revealing is too long or too complex to be coded in its entirety; then the unit of analysis will be broken down into smaller units of coding. Again, in a typical study, the so-called boundaries of a unit of analysis and a unit of coding are the same. For example, in a study where the construct of interest is personality, the person represents the unit of analysis and the unit of coding. The unit of context sets additional contextual boundaries. Suppose, for example, a researcher wants to examine stereotypes in TV commercials. Each commercial is typically a unit of analysis, which will need to be broken down into smaller units of coding to be able to capture instances of stereotypes in every commercial. The unit of coding could be the central character of a commercial who either behaves in a stereotypical or a nonstereotypical way. The unit of context will further specify how long a researcher has to observe that individual and/or in what context. The two concepts, the unit of coding and the unit of context seem similar, and the best way to separate the two is to say that what the research will ultimately code is the *unit of coding*. The information about the unit of analysis, the unit of coding, and the unit of context is usually included in *a coding scheme*,

along with the details/instructions about how the construct of interest is going to be operationalized and coded.

e. Oftentimes, the categories that describe a phenomenon of interest must be assigned numeric codes (e.g., from 1 to 10) to enable a researcher to analyze the data. If the codes are only meant to distinguish categories, as in "1" for a stereotypical portrayal of a female character and "2" for a nonstereotypical portrayal, then the scale of measure of this variable is nominal. If the categories are meant to represent some type of ranking and the variable is thus measured on an ordinal scale, the codes should be assigned accordingly. For example, if the categories are *highly stereotypical*, *stereotypical*, and *not stereotypical*, then the codes should reflect the degree from high to low, as in "2" for highly stereotypical, "1" for stereotypical, and "0" for not stereotypical. Finally, if a researcher is only interested in measuring the number of times female characters were represented in a stereotypical way, then no codes are necessary; the researcher will simply count the instances of the stereotypical representation. In this case, the variable is measured on a ratio scale, as the measure of stereotypical representation represents a true, meaningful number.

f. Once the construct is coded, this quantitative data can be transferred to SPSS, a statistical software package, (or to any other statistical software package of your choice) to be analyzed. Typically, the first column is reserved for the unit of analysis (e.g., the people, commercials, schools, or categories) with the scores, if a study involves categorizing and counting the number of observations that fall within each category. The subsequent columns represent the variables or measured constructs.

    - For example, in a typical study where each participant is measured on some dimension of personality, the first column will represent the participants, while the second column will have the numerical measure of personality corresponding to each individual. In content analysis, the first column could be commercials. For example, take a look at Figure 1, displaying a hypothetical data set in SPSS. The number one in the first row and the first column, labeled as *Commercials*, represents commercial number one (it can also be given a name), and the score of "3" in the same row and in the column labeled *Stereotypes* (e.g., the frequency of a stereotypical portrayal) is a numerical measure of stereotype, supposedly found in the first commercial. Notice that the scale of measure in this example is a ratio, as each number represents a true count of stereotypical representations of female characters.



	Commercials	Stereotypes
1	1	3
2	2	4
3	3	5
4	4	3
5	5	4
6	6	5
7	7	4
8	8	5
9	9	3
10	10	3
11	11	5
12	12	4
13	13	5
14	14	5

Figure 1. *Example of coded data entered in SPSS.*

Next, we will take a look at the four studies mentioned above to illustrate the elements of content analysis.

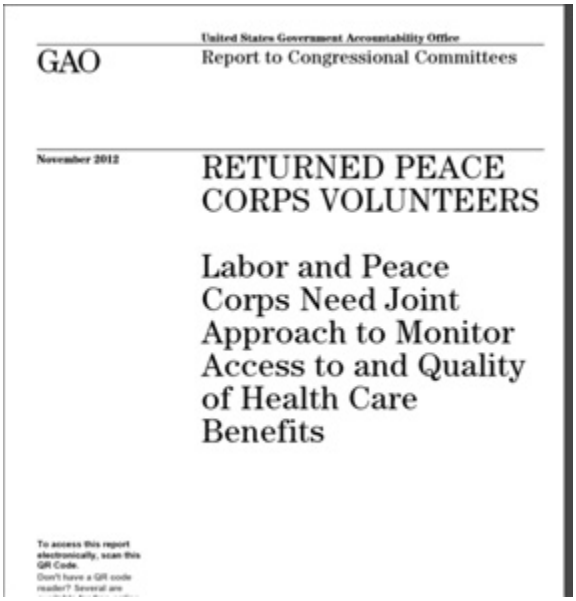
4.2 Examples

Study 1

Program evaluation has several elements of a scientific study. For example, just as in research, a program analyst sets the goals or the questions to be answered (similar to research hypotheses) and the plan for how they will be answered or executed (similar to a research design). However, unlike in a scientific study, where no judgments should be placed, the goal of a program evaluation is "to determine the merit, worth or value" of a program and to give evaluative feedback (Coffman, 2003). Still, a researcher or a program analyst can use similar methods, such as content analysis, to test a hypothesis or to determine the quality of a program.

To fulfill my MPA (Master of Public Administration) degree requirements, I worked as an intern in the US Government Accountability Office. One of my assignments was to work in a team that

was evaluating the quality of accessibility of health care benefits to Peace Corps volunteers by the Department of Labor and the Peace Corps.[3]


United States Government Accountability Office

GAO

Report to Congressional Committees

November 2012

RETURNED PEACE CORPS VOLUNTEERS

Labor and Peace Corps Need Joint Approach to Monitor Access to and Quality of Health Care Benefits

To access this report electronically, scan this QR Code.

Figure 2. *Front page of the* Returned Peace Corps Volunteers *report.*

One of the methods used to evaluate the quality of the program was the content analysis of the letters that were sent to inform the applicants about the denial of their health care coverage applications. The goal of this analysis was to find out if there was a breakdown at the management and administrative levels of the health care system that led to the denial of the health care coverage. One of the steps to determine that was to find out what the reasons were for denial of medical coverage and how frequently they occurred.

a. The ***overarching construct*** was the *quality* of health care benefits provided by Labor and Peace Corps to returned Peace Corps volunteers; one of the operational definitions for the *quality* of health care benefits was *accessibility* of health care for qualified Peace Corps volunteers. The denial of accessibility was, of course, the opposite of the quality of health care.
b. The ***unit of analysis*** was every denial letter. But we can also call the reasons for denial of health care the unit of analysis, as the conclusions were made about the reasons for the denial of coverage described in the letters.
c. The denial letters served as data and were used to determine the reasons why some Peace Corps volunteers were denied coverage or access to health care benefits. Thus, the ***unit of coding*** was every denial letter. The letters did not follow any particular format, so, in order to determine the reasons for denial, an entire letter had to be evaluated. The reasons for

[3] *To obtain the full copy of the GAO-13-27 report, go to gao.gov or contact Linda T. Kohn at (202) 512-7114 or kohnl@gao.gov.*

denial were then classified into several broad categories, for example, *failure to include a medical diagnosis*, *failure to provide connection of illness to Peace Corps services*, and others. In order to compute any statistical analysis, every category (i.e., denial reason) was assigned a code. Suppose there were a total of 5 reasons; each reason was assigned a code from 1 to 5 to simply distinguish them rather than to assign them some degree of denial. In other words, denial due to failure to include medical diagnosis, assigned as "1," is no worse or better than denial due to failure to provide connection of illness to Peace Corps services. Thus, the measure of scale, in this example, is nominal.

This is what the data could have looked like:

Unit of analysis	Variable (Reasons for denial)
Denial letter #1	2 (i.e., failure to provide connection…)
Denial letter #2	1 (i.e., failure to include medical diagnosis)
Denial letter #3	5

Study 2



The purpose of the study conducted by McAdams, Diamond, Mansfield, and Aubin (1997) was to test whether, and in what way, more generative and less generative older adults differ in how they identify themselves as individuals in the past, present, and future by asking them to create a narrative of their lives. The idea of generativity comes from Erikson (1950) and his developmental theory of personality. He essentially theorized that in middle and late adulthood people begin to face with a critical task of either starting to invest their lives in the future of the next generation (i.e., caring about what they leave behind as their legacy) or focusing more on oneself. The latter, Erikson believed would lead to stagnation, while the former, generativity (i.e., conceptually defined as "having concern for and commitment to the well-being of the next generation", Erikson, 1963) would bring life satisfaction. These ideas were tested empirically in various studies. A good example of one such research is by McAdams and colleagues (1997). Going back to their original research question, they hypothesized, based on the extensively reviewed literature, that there would be differences in the stories of *commitment* of more and less generative adults. In essence, commitment themes can be said are the way the authors operationally defined generativity. However, commitment themes were further operationalized into five specific types (see below). To test it, they collected a sample of 70 and older adults, asking them to narrate their life story.

Below are my examples of what I think the components of their content analyses could be:

a. We can say that **generativity** is the main construct which was operationally defined by stories of **commitment**. Based on theoretical and empirical evidence, the authors further operationally defined commitment themes into the following categories or themes: *(1)early advantage, (2) sensitivity to the suffering of others, (3) moral steadfastness, (4) redemptive sequences of life scenes,* and *(5) commitment to prosocial goals*. The narratives relating to all five themes were elicited from the participants by asking them semi-structured questions. Furthermore, each category was also operationally defined to help the coders determine the presence of each theme within the narratives. For example, one of the indicators that operationally defined early advantage was *family blessing* (i.e., when the narrator described or referred to himself/herself as being blessed for having some special characteristics or qualities in his/her childhood).
b. The ***unit of analysis*** was each participant.
c. The ***unit of coding*** could be an entire interview or it could be portions of the participants' responses. The coder would have to listen to the entire or a portion of the interview with each participant to identify themes of commitment. The instances of the themes defining each category of commitment story most likely were later calculated for each participant and used as variables for further statistical analysis.

This is what the data could have looked like:

Unit of analysis	Commitment (Variables)						
	Family Blessing theme (early advantage)	Childhood Attachment (early advantage)	Helpers (early advantage)	Suffering of Others	Moral Steadfastness	Redemption Sequences	Prosocial Goals
Participant #1	3	2	2	3	3	2	3
Participant #2	2	1	2	1	0	3	1
Participant #3	1	0	0	2	1	1	2

Study 3



The study of TV commercials (archival data) by Mastro and Stern (2003) is a good example of a content analysis study with more than one unit of analysis. The purpose of the study was to learn more about the frequency of racial and ethnic minority characters being portrayed in TV commercials, as well as the context of those portrayals. They also tested whether there were any differences in the frequency and/or context of racial and ethnic minority characters and white characters being depicted. The authors used a cognitive learning perspective to argue that TV viewers, especially children, tend to identify themselves more with images of people of their own racial and ethnic background. The authors argued that when TV viewers are frequently exposed to distorted or stereotypical images of people of their own racial or ethnic background, they are more likely to internalize them and develop unhealthy self-perceptions. This only further reinforces stereotypes and biases. In addition to the characters and the contexts in which they appeared, the study kept a record of the products that each commercial advertised, to find out if there were any systematic differences in the products that black and white characters advertised. The sample of the study consisted of 2,880 commercials drawn from prime-time television programming across six major networks: ABC, CBS, NBC, FOX, UPN, and WB.

a. The study had five specific hypotheses and multiple ***constructs*** (these were constructs that would provide information about the differences in the contexts in which commercial characters appeared). These were *product type*, *setting*, characters' *relationship to the*

product (using, endorsing, both, or neither), *behavior*, *job authority*, *social authority*, *family status*, *alluring behavior*, *sexual gazing*, *degree of dress*, *hierarchy position*, *extent to which characters are respected*, *activity, physical attractiveness*, *body type*, and *age*. Each construct had to be ***operationally defined***. Here are just a few operational definitions that the researchers used to measure their constructs:[4]

- Setting is operationally defined as *work*, *home*, *other indoors*, and *outdoors.*
- Relationship to the product is operationally defined as *using*, *endorsing*, *both*, or *neither.*
- Degree of dress is operationally defined as the attire of the character ranging from *conservative* (coded as 1) to *suggestive* (coded as 5).
- Sexual gazing is operationally defined in terms of 4 possible options—*receiving a sexual gaze*, *giving a sexual gaze*, *both*, or *neither.*

In order to conduct any statistical analyses, the researchers would have also assigned numerical values to each category of the constructs. These numbers would have been nominal, as their only purpose was to distinguish the categories (e.g., work setting, home setting, other indoor/outdoor setting). For example, work setting could be assigned "1"; home setting, "2"; and so on.

b. ***The unit of coding*:** The first unit of coding was an entire commercial, as the product type variable had to be measured at the commercial level (in other words, a coder would have to watch an entire commercial to understand what it advertised). All other constructs had to be captured at the character level— thus, the character (his/her behavior, setting, etc.) was the second unit of coding.
c. ***The unit of analysis*:** The study was interested in both the character representations and the kinds of products that each commercial advertised; thus, there were two units of analysis: commercials and commercial characters.

This is what the data could have looked like:

Unit of analysis		Variable #1	Variable #2	Variable #3
Commercials	Character	Product	Setting	Relationship
Commercial #1	White male	1 (financial services)	1 (work)	1 (using)
	White female		1 (work)	2 (endorsing)
Commercial #2	Asian male	2 (technology)	1 (work)	1 (using)

[4] *The complete list of the operational definitions can be found in the article.*

Study 4

Content analysis has useful applications for evaluating observational data and converting images, sounds, or actions into numbers. A good illustration of this is the study from my dissertation. Specifically, I was interested in studying intersubjectivity between mothers and their two-and-a-half-year-old toddlers. To measure this construct, 79 dyads (mothers and their children) individually were invited to our lab and asked to read a couple of books for about 10 minutes. Seventy-nine mother-child dyadic book-reading activities were separately recorded and later analyzed using content analysis.

a. ***Construct*** to be transformed into a variable: Nonverbal intersubjectivity, conceptually defined as shared understanding. Drawing on insights from several theories and prior investigations, I ***operationally*** defined nonverbal intersubjectivity as *affective synchrony* and *affective matching* between mother and child—a combination of facial expressions and gazing that both the mother and child shared during book reading.
b. ***Unit of analysis*****:** Each mother-child dyad—my analyses encompassed two people, the mother and the child, rather than each separately. In other words, I was making conclusions about the significance of intersubjectivity between a mother and a child (a dyad) rather than a mother or a child as a separate unit. I found that the harmony between the mother and the child (i.e., as a dyad) seemed to play a more important and positive role on the child's emotional development than the quality of the maternal or the child's behavior alone.
c. The ***unit of coding***: In order to capture dyadic intersubjectivity, I broke down the average of 10-minute book-reading interaction of each dyad into 5-second intervals. This was necessary to capture the moment-to-moment variations in the mothers' and children's facial expressions and gazing, as well as to make the coding more manageable. At the end, I calculated the proportions of time the mother and their children were in sync in their expression of affect and eye contact to represent the degree of nonverbal intersubjectivity.

Here is what the data could have looked like:

Unit of analysis	Variable # 1	Variable #2
	Affective synchrony	Gaze synchrony
Mother-child dyad #1	100/120	50/120
Mother-child dyad #2	10/120	3/120

4.3 Content Analysis of Language

The logic and the principles of content analysis can be applied to spoken data (e.g., to a conversation between friends or coworkers, etc.). The goal is the same, to find any systematic and meaningful variations or patterns in the content or structure, which can be possibly used to predict future behavior.

For example, in my earlier example of the study on mother-child intersubjectivity, the goal was to measure the quality of mother-child book-reading nonverbal interaction by looking at the mother and the child as one unit. Similarly, we can also evaluate the quality of their verbal interactions, for example, to see how connected or in sync they become the longer they converse. Researchers Ensor and Hughes (2008) examined something similar to that idea. Specifically, they analyzed conversations between mothers and their two-year-old children to test a hypothesis that the degree of connectedness between mothers and their young children would contribute to children's development of social understanding (e.g., understanding about the social rules and conventiosn) more strongly than the actual words that the mothers used in their conversations with their children.

a. The authors operationally defined *connectedness* as the following: when a speaker's utterance is semantically related to the other speaker's previous utterances.
b. Each mother-child conversation was segmented into turns or *units of coding* that were defined as "utterances of one speaker bounded by another speaker's utterances or a significant silence (usually 5 s or more)" (Shatz and Gelman, 1973).Each turn was categorized as *Connected*, *Initiation*, *Failed*, *Conflict*, or *Unclear* (all categories were operationally defined as well). Of course, the *Connected* turns category was the main focus of the study.

Here is what the data could have looked like:

Unit of analysis	Connected turn	Initiation turn	Failed turn	Conflict turn	Unclear turn
Mother #1	50	25	30	10	2
Child #1	20	5	25	50	20
Mother #2	20	10	30	25	2
Child #2	5	0	50	20	15

4.4 Inter-Rater Reliability

Since content analysis relies heavily on the accuracy of coding, establishing reliable coding scheme is a major element of this design. Reliability refers to the idea of consistency of a measure; whereas, inter-rater reliability refers to the consistency in the coding between two or more independent coders. To make this concept even broader, reliability speaks to the replication of the measure. If consistency is high, then the researcher can be confident that the data is reliable (i.e., replicable). Typically, inter-rater reliability is computed during the developmental phase of a coding scheme. Specifically, two or more coders are trained to code about 15% to 20% of the same data in order to check the reliability of the coding scheme. If the analysis shows that the agreement between the coders is high enough, the coding scheme is viewed reliable and the coding can proceed. If the agreement is too low, the coding scheme must be revised until acceptable agreement between the coders can be established.

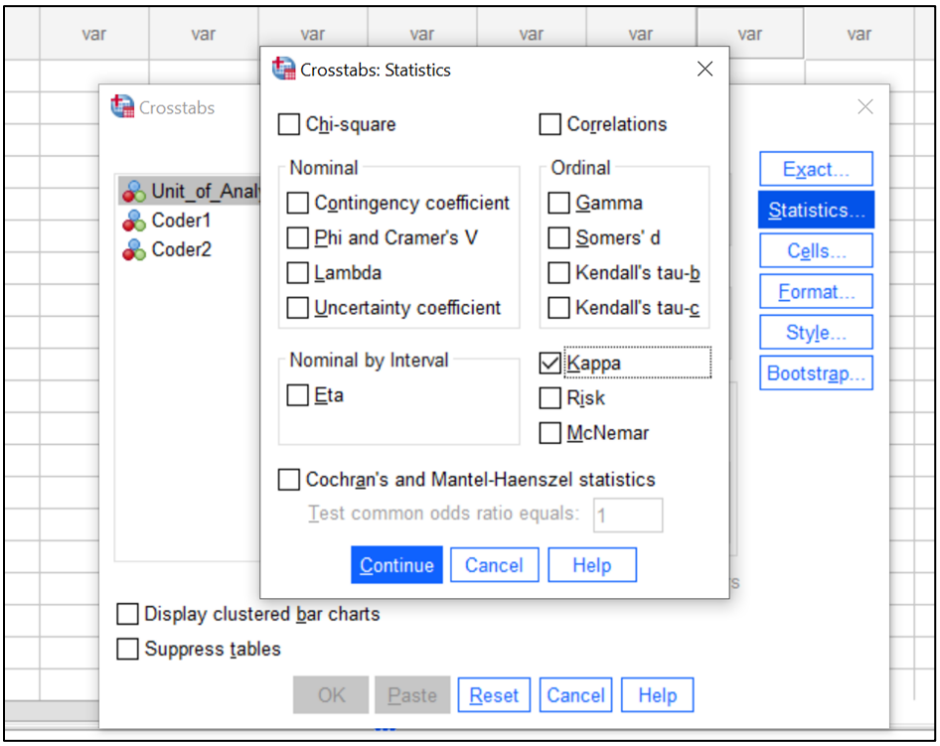
Two most commonly used inter-rater reliability statistical tests are Cohen's kappa and intra-class correlations. The former is used when the variables are nominal; and the latter should be used if the data are ordinal, interval or ratio.

Computing Cohen's Kappa

Suppose two coders coded independently the same data using some nominal scale coding scheme, '1' and '0' (see example below).

Unit_of_Analysis	Coder1	Coder2
1	1.00	1.00
2	.00	.00
3	1.00	1.00
4	.00	.00
5	.00	.00
6	.00	1.00
7	1.00	1.00
8	.00	.00
9	.00	1.00
10	.00	.00
11	.00	.00
12	.00	1.00
13	.00	1.00
14	.00	.00
15	1.00	1.00
16	1.00	.00

To compute Cohen's kappa, follow these steps in SPSS: go to Analyze >◊Descriptive statistics > Crosstabs > move one Coder to Column(s) and another to Row(s)◊> click on Statistics > click Kappa



The most important table of the test is 'Symmetric Measures', which includes the measure of agreement (see example below).

Coder1 * Coder2 Crosstabulation

Count

		Coder2		
		.00	1.00	Total
Coder1	.00	7	4	11
	1.00	1	4	5
Total		8	8	16

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.375	.215	1.618	.106
N of Valid Cases		16			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

The rule of thumb is considering values 0.0 to 0.2 as no-to weak agreement, 0.21 to 0.40 fair agreement, 0.41 to 0.60 moderate agreement, 0.61 to 0.80 strong agreement, and 0.81 to 1.0 almost perfect or perfect agreement (Hallgren, 2012). Thus, the agreement in my example can be considered only fair; and a researcher would have to decide whether such agreement is acceptable to proceed with the coding scheme. Of course, the issue may not always be faulty coding scheme—for example, it is also possible that the coders need additional training to increase the reliability and accuracy of their work.

Computing Intra-class Correlations (ICCs)

As I mentioned earlier, if the data that you are checking for inter-rater agreement is either ordinal, interval or ratio, the appropriate statistics is ICCs.

Suppose, two coders were to count occurrences of some phenomenon per unit of analysis (see example below).

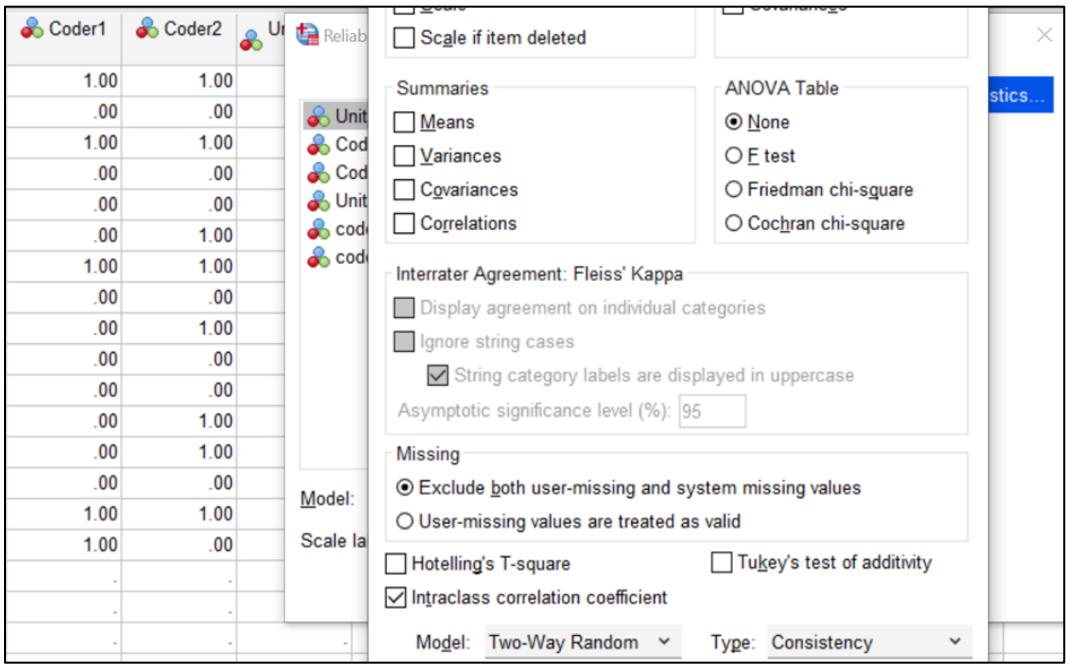
UnitofAnalysis	Coder_1	Coder_2
1	1.00	1.00
2	4.00	1.00
3	2.00	2.00
4	3.00	3.00
5	3.00	3.00
6	1.00	3.00
7	1.00	2.00
8	1.00	1.00
9	1.00	1.00
10	1.00	1.00
11	3.00	2.00
12	2.00	2.00
13	1.00	2.00
14	3.00	3.00
15	3.00	3.00
16	1.00	1.00

To compute how closely in agreement the two coders are, start by going to Analyze◊Scale◊Reliability Analysis◊move the coders to Items◊click on Statistics◊click on Intraclass correlation coefficient…STOP here for now and consider your 'Model' and 'Type'.

First, under Model, there are three different models you have to choose from. Model 1 'Two-Way Mixed' should be chosen if you are only interested in measuring reliability of the two selected coders. In other words, the measure of agreement will not be generalized to other potential coders. Model 2 'Two-Way Random' is the model of choice if the measure would be generalized to other potential coders with the same characteristics. Finally, model 3 'One-Way Random' should be chosen if the phenomenon is coded by a set of different coders. This last scenario would be an unusual one, however, in some circumstances this may happen (Koo & Li, 2016).

Under 'Type', you have two options, 'consistency' or 'absolute agreement'. The former concerns with the degree to which the coders correlate cumulatively, whereas the latter refers to the exact agreement between the coders on the same set of score(s) (Koo & Li, 2016).

Suppose, I am interested in a Two-Way Random, and consistency type agreement:



The most important table that the SPSS will produce is the Intraclass Correlation Coefficient table (see below):

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.426[a]	-.071	.753	2.484	15	15	.044
Average Measures	.597	-.152	.859	2.484	15	15	.044

Two-way random effects model where both people effects and measures effects are random.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency definition. The between-measure variance is excluded from the denominator variance.

You will find two measures: single and average. The choice depends on whether you/the researcher will intend to use the average of the codes of the two coders as your final data, or whether you will use codes of only one of the coders. If the former, then the 'average' measure should be used to interpret inter-rater reliability; however, if it is the latter, you must use the 'single' measure (Koo & Li, 2016).

The rule of thumb for interpreting the magnitude of the ICC values is as follows: any values less than 0.5 indicate poor reliability; values between 0.5 and 0.75 indicate moderate reliability; and anything above 0.75 is good to excellent (Koo & Li, 2016)..

So depending on how I intend to use my coders, the reliability in my example can be considered low (if I have to use the single measure) or acceptable (If I use the average measure).

Video Example of a Content Analysis of Video Games

https://sentia.matrixlms.com/files/1008444/Senior_Thesis_Presentation_-_Google_Slides_-_Google_Chrome_2021-01-27_20-40-43.mp4?lmsauth=2527d8eb3210227f449a4029574bd89806fe311e

Guided Practice (not graded)

To test your understanding of the process involved in content analysis and the logic behind it, develop a brief coding scheme by identifying a unit of sample, a unit of analysis, and a unit of coding for the following hypothetical study.

An instructor would like to compare five written essay assignments to find out which assignment demands students to engage more frequently in critical thinking. She operationally defines critical thinking as "several sentences that challenge the assumptions of the course with well-reasoned and knowledge-based argument(s)." The same 50 students will write each of the five assignments, and the frequency of critical thinking in each assignment will be counted to provide the measure of the demand for critical thinking.

a. What is the construct to be measured?
b. What is the unit of sample?
c. What is the unit of analysis?
d. What is the unit of coding?

Answers to Guided Practice Exercise

Based on the details provided above about the data collection, the instructor will most likely intend to study the phenomenon in the following way:

a. The construct of interest is *critical-thinking skills.*
b. The unit of sample is *students and their five papers.*
c. The unit of analysis is *students, as they will provide measures of construct*. The easiest way to understand this step is to visualize how the data will be entered. See table below:

Unit of analysis	Paper #1	Paper #2	Paper #3	Paper #4	Paper #5

Student #1	10 (frequency of critical thinking)	5	10	15	25
Student #2	15	4	5	15	20
…Student #50	20	6	4	10	15

d. The unit of coding can be a paragraph. In other words, each paragraph will be evaluated for *thematically connected sentences*, to identify and count frequency of critical thinking per each paper.

According to the results (look at the frequency counts of critical thinking sentences in each column), paper #5 demands the most critical-thinking skills from the students.

References

Beebe, B., & Lachmann, F. (1988). Mother-infant mutual influence and precursors of psychic structure. In A. Goldberg (Ed.), *Frontiers in self psychology* (Vol.3, pp. 3-25). Hillsdale, NJ: The Analytic Press.

Bengtsson, M. (2016). How to plan and perform a qualitative study using content analysis. *Nursing Plus Open, 2*, 8-14.https://doi.org/10.1016/j.npls.2016.01.001

Coffman, J. (2003). Ask the expert: Michael Scriven on the Differences Between Evaluation and Social Science Research. *The Evaluation Exchange*, *9*(4). Retrieved January 8, 2012 from http://www.hfrp.org/evaluation/the-evaluation-exchange/issue-archive/reflecting-on-the-past-and-future-of-evaluation/michael-scriven-on-the-differences-between-evaluation-and-social-science-research.

Erikson, E. H. (1963). *Childhood and society* (2nd ed.). New York: Norton.

Gerhards, J. (2010). Non-discrimination towards Homosexuality: The European Union's policy and citizens' attitudes towards Homosexuality in 27 European countries. *International Sociology, 25*, 1-28.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutor Quant Methods Psychol, 8*, 23-34.

Holsti, O. R. (1969). Content analysis for the social sciences and humanities. Reading, MA: Addision--Wesley.

Hoy, W. K. (2009). *Quantitative research in education: A primer*. Sage Publications, Inc.

Klimenko, M. A. (2012). *Mother-toddler inter-subjectivity as a contributor to emotion understanding at age 4* (Unpublished doctoral dissertation). University of Georgia, Athens.

Koo, T., K., and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability. *Journal of Chiropractic Medicine, 15*, 155-163.

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3rd ed.). Sage Publications, Inc.

Mastro, D. E., & Stern, S. R. (2003). Representations of race in television commercials: A content analysis of prime-time advertising. *Journal of Broadcasting and Electronic Media*, 638-647.

McAdams, D. P., Diamond, A., Mansfield, E., & Aubin, E. (1997). Stories of commitment: The psychosocial construction of generative lives. *Journal of Personality and Social Psychology, 72*, 678-694.

Plaut, V. C., Markus, H. R., Treadway, J. R., & Fu, A. S. (2012). The cultural construction of self and well-being: A tale of two cities. *Personality and Social Psychology Bulletin, 38(12)*, 1644-1658.

Tse, D. K., Belk, R. W., & Zhou, N. (1989). Becoming a consumer society: A longitudinal and cross-cultural content analysis of print ads from Hong Kong, the People's Republic of China, and Taiwan. *Journal of Consumer Research, 15*, 4, 457-472.

U.S. Government Accountability Office. (2012). *Returned Peace Corps volunteers* (GAO-13-27).

Chapter 5: Observational Method

Introduction

In this chapter, we will continue covering non-experimental research methods, and will focus, specifically, on the observational method. This is one of the most commonly used research methods in the field of psychology and social sciences.

5.1 Reasons to Use Observational Method

Observational design falls within a non-experimental category of research methods. This means that unlike an experiment, and similar to a content analysis, observational studies cannot test causal relationships between variables. Thus, a researcher, who hypothesizes that playing violent video games causes aggression, cannot use an observational method to test such a hypothesis. However, this doesn't mean that an observational method isn't valuable. On the contrary, it is one of the most widely utilized methods in psychological research. In fact, this method has been called "the necessary link between laboratory research and "real-world" behavior, and the bane of our aspirations for more accurate, more objective information about behavior." (Altmann, 1974). The reasons and advantages of gathering data with an observational method are discussed next.

When the Research Goal is to Describe

Recall that one of the goals of scientific research is to describe. This goal is easy to underestimate because, after all, only discovering what causes an event to happen is the kind of knowledge that explains why it happens and can lead to better prediction and control. While a scientifically derived description of an event or behavior can still provide enough information to predict and even control the behavior, we may never have a full grasp of what is being described if we don't know the underlying causal mechanisms. Still, despite the limitations of descriptive knowledge, it can yield very rich data that can be used as the first step towards understanding the phenomenon of interest. For example, a researcher who wants to test a hypothesis that playing violent video games causes aggression may first want to know about different ways that aggression is manifested when someone plays violent video games. Thus, a study, detailing people's behavior as it relates to the frequencies or kinds of video games played, can be an incredibly useful one.

An Example of Research on Inter-Subjectivity

To further illustrate the point that observational research can bring about new knowledge and stimulate more exciting research, I will proffer a very brief synopsis of the research on *inter-*

subjectivity, a construct that is loosely defined as, shared understanding between people. In the 1970s, social and developmental psychologists worked on an interesting phenomenon using mainly observational methods—the tendency for two (or more) interacting individuals to synchronize their gaze, posture, affect, and gestures with each other. The rhythmic patterns and mimicking in human social interactions were observed not only between adults but were also witnessed between caregivers and the caregivers' newborns, which was pointing to some biological foundation supporting this observed synchrony as early as infancy. Using a variety of different observational methods to measure synchrony, by calculating the proportions of matched facial expressions or mutual gazes, developmental psychologists were able to determine that infants who routinely had a higher affective or gaze synchrony with their caregivers had better developmental outcomes. These children were better at regulating frustration, had more advanced language skills, and had a better understanding of other people's perspectives and emotions.

These studies led to several theories and many hypotheses trying to explain the developmental and social significance of the human tendency to synchronize. Using experimental methods and with the help of more advanced technologies that were developed over the years, researchers have begun to learn more about the biological underpinnings of interactional synchrony. One of the interesting discoveries in this area of study was made by Vittorio Gallese in his experimentations with macaque monkeys, the discovery of the so called, 'mirror neurons'(Gallese, Fadiga, Fogassi, & Rizzolatti, 1996; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996). We now know that mirror neurons also exist in humans. What is unique about these sets of neurons is that they are activated either when an individual performs an action or when an individual observes another person performing the same action--. mirroring the same or similar sensation that the observed individual is experiencing. Many researchers believe that this can explain how we, as humans, can relate to one another on an intimately human level. This may explain why we interact almost in sync with another person, and how we can 'feel' another person's emotional or physiologically sensations (e.g., Gallese, 2002). This phenomenon came to be known as inter-subjectivity.

An Example of Research on Infant Cry

Knowing something in terms of its descriptive qualities can be a goal in itself, but, no less valuable. For example, Zeskind and Collins (1987), conducted an observational study in a day-care center to learn more about infant's cries during the first 35 months of life. They also wanted to observe how caregivers tended to respond to them. Researchers taped the cries and observed the adults' behavior as they responded, if at all, to the cries through a one-way window. Observers noted the types of responses the adults made, e.g., whether they picked up the baby, or ignored the cry, etc., and also rated the degree of the perceived urgency with which the adults would respond to the cries. What they learned was that the cries varied a great deal in terms of both their length, ranging between 18 and 577 seconds, and their acoustic features. Infants would usually emit a high-pitched cry when they were in pain, separated from their parents, or when their food was removed. Researchers also confirmed what previously had been found, that high-pitched cries were most

urgently responded to, and that adults were more likely to pick up and rock those babies whose cries were high-pitched rather than low-pitched.

There are several practical and theoretical benefits to knowing this information. For example, knowing that frequent and, sometimes, prolonged infant crying is a normal and expected phenomenon, first time parents can better prepare themselves to deal with the psychological and physical stress that is, inevitably, associated with it. Knowing this can also help reduce the frustration and the guilt that some parents may feel when they are unable to sooth their crying baby. Finally, these studies have shed light on the functional importance of infant crying. The research suggests that it serves as a communication signal to let the caregiver know that the infant is in distress and needs physical or emotional support.

To Complement Other Research

Recall that while experimental research is ideal if you want to achieve the maximum level of control and internal validity, the subjects or participants in these studies are usually tested in artificially created and controlled settings and are exposed to artificially manipulated, independent variables. It is also not uncommon for researchers to conduct experiments using animals in their inquiries about human behavior and processes. Researchers can control and manipulate situations with the animals that they would not be normally able to do with people for moral or ethical reasons. For example, researchers can use invasive techniques such as directly attaching recording devices to the brains of the animals, removing some parts of the brain, or cutting connections between brain areas. Research on posttraumatic stress disorder (PTSD), characterised by "hyper-arousal, disturbing flashbacks and numbing or avoidance of memories of an event" (American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5. 5th ed.) commonly relies on animal models, for example exposing rats to various stressors and observe the direct effects of that exposure: e.g., electrical footshock stress, physical restraint for an extended period of time, etc.

While these are, no doubt, very important studies, they tend to have weak external validity, especially when the studies use animals to study human behavior and mental processes. For example, how comfortable will you feel about taking a new drug for depression if you knew that its effectiveness had only been studied and supported in experiments with animals? Surely, these experiments would have high internal validity (i.e., the environment was highly controlled to detect the effect of the manipulated variable), but the drug effectiveness would most certainly have to be studied using real people before its effect can be confirmed and safe to give to people (i.e., can be generalized to the general public of people battling depression).

So, one way to address the low external validity of experimental research is to complement its results with those derived from observational studies. For example, observational research can complement experimental and other types of research by providing details of the settings and the

dynamics in which certain behavior occurs (e.g., research on intervention in bullying on school playgrounds). Other times, observational studies supplement self-reported measures to make sure that they are accurate and unbiased.

An Example of Research on Peer Intervention in Bullying

Imagine you are studying bullying in school and would like to know not only who does the bullying, but also how it happens, what leads up to it and how the victim, and the others witnessing it, behave towards the bully and the victim. All of these are important questions that will not only help school administrators and parents better understand the causes of bullying but also how to address it. So, what would be your preferred method of data collection? It calls for a study that would use an observational method.. You would want to have a realistic and detailed description of the people involved in the bullying including the bully, the victim, and the witnesses and the actions that they take either to stop or continue it. So, researchers, Hawkins, Pepler and Craig (2001), conducted an observational study to answer the following questions: (1) how frequent bullying occurs, (2) how often and what actions the peers who witness it take to intervene and, (3) how effective their interventions are. The last question was particularly important to address because peer-intervention may be one of the effective ways to stop or even prevent bullying. However, some experimental studies have not been able to support the idea that peer-intervention can lead to a decrease of bullying. Thus, it was important for the researchers to conduct studies that would observe children in their natural settings and behavior to be more assured of the effectiveness of peer-interventions to stop bullying.

An Example of Research on Drinking

Researchers Jeremy Northcote and Michael Livingston conducted an observational study in 2011 to examine the accuracy of self-reports of drinking. Specifically, they wanted to know how accurate and honest people self-report their own frequencies of drinking in surveys. To do that, the researchers recruited 11 undergraduate students to follow and observe a total of 62 students during various drinking sessions; the students gave permission to participate and to be observed for the study by the undergraduate research assistants. Afterwords, the observed students were asked about the amount of alcohol they remembered consuming during the observed sessions; and their reports were compared with the recorded data. The results revealed that the self-reports were fairly accurate when the subjects engaged in light to moderate drinking; however, when the drinking was heavy (more than eight drinks per session) the self-reports were underestimated.

Take-away Message: Although I grouped the aforementioned studies into two categories (when the goal is to describe or to complement other research), rarely researchers think in terms of one or the other specific goal --they choose an observational method for all of its benefits and despite the drawbacks because ***it is compatible with your research goal***. For example, the study by

Hawkins and colleagues (2001) conducted the study to better understand the natural context in which interventions to bullying occurs or fail to occur. Their findings not only offer an important insight into the natural interactions between the school bullies, the victims, and the by-standers, but also supplement the studies of children's attitudes toward bullying using self-report measures.

5.2 Main Features of an Observational Method

Types

Observational studies can take place in natural settings, where the behavior or the event that is observed naturally occurs. This type of observational method is called naturalistic. The study on bullying, by Hawkins, Pepler and Craig (2001), is an example of a naturalistic, observational study, where bulling was observed on the school playground—one of its natural environments. The main and unique feature of the naturalistic approach, and its likely advantage, is that there is minimum interference from a researcher which preserves the natural dynamic of what is being observed. For example, the only interference from the researchers that took place in the bullying study was that the participating children had to wear a microphone and a waist pouch with a wireless transmitter so that their interactions on the school playground could be recorded.

Researchers can also structure observations to have more control over the observed phenomenon, the environment, the participants, and their behavior. These *structured* observations will usually take place in a lab where extraneous variables are easier to control; but the lab can also be set up in such a way as to resemble a more natural environment. For example, the lab can be set up to look like a living room in order to make participants feel more at ease or to serve some other purpose. In addition to observing participants in a laboratory, researchers may have scripted communication with the participants, instructing them on the tasks and how they are expected to perform them. This, and any other standardization of a structured observational study, will eliminate the possibility that the differences in the administration of the study, e.g., the style of a researcher's communication, and not the actual predictor, contributed to the outcome.

For example, if the purpose of a study is to find out whether the personal characteristics of a mother and her child might affect the dynamic between the two during a book reading interaction, ideally, a researcher would like to eliminate any *extraneous* environmental factors that could also affect their behavior during the book reading. This is done in order to be certain that any observed variations in their behaviors are likely to be due to their individual characteristics only. To achieve control over environmental differences such as, book choices or home environment, it would be best to observe every mother-child dyad reading the same book during the same period of time after receiving the same instructions in a standardized laboratory setting. This should be accomplished with professionally set up cameras, capturing the best view of the participants' interactions. However, if the purpose of the study were to find out what books mothers and

preschoolers tend to read at home the most, a naturalistic approach, without interference from a researcher, would be a more appropriate observational method.

Of course, there is an ongoing debate as to which approach is more desirable. For example, one one hand, one can argue that structured observations elicit the behavior of interest more efficiently, and enables the comparisons between individuals to be more valid (since they are compared across individuals performing the same task and in the same environment). On the other hand, some researchers have expressed concern over the validity of the measures derived in artificial settings, like in a laboratory, where individuals are asked to perform certain tasks rather than are allowed to do what they normally do at home or in their natural environment.

Important Concluding Remarks: In the end, the choice of whether to use a naturalistic or a structured laboratory observational method will depend on the goals of the research. Generally, if the main goal is to capture a phenomenon of interest in its natural state, then a naturalistic approach is the best method for the task. For example, if a researcher's specific research goal is to study workplace deviance, it would seem more important to conduct the study in the actual workplace setting rather than to be able to control extraneous environmental factors. Thus, it would warrant a naturalistic observation. Alternatively, if the purpose of the study is such that it requires having control over the environment in which a phenomenon is being observed, and if elicitation of some specific behavior is necessary for comparison purposes, then a structured laboratory observational approach is a more appropriate choice. For example, if a researcher wants to compare the interactions during book reading session mothers and children of high- and low-socioeconomic backgrounds, then a structured observational method is more desirable as it will allow the researcher to standardize the environment (to control for potential differences in living conditions, books availability) and to focus specifically on the differences in the behavior of the dyads.

Transforming Constructs into Study Variables

Methods of Quantifying Behavior and Measuring Scale

Recall that quantitative research has to generate numeric data to test hypotheses and to make inferences. Thus, all observations have to be converted into numbers—the constructs have to be transformed into variables. The most common methods of quantifying observations are:

- Frequency (of an event)—by counting the number of occurrences of the behavior or the construct of interest. Typically, an event is the type of behavior that occurs instantly, for example, a smile or a hit (Altmann, 1974). Important to distinguish between raw count and a rate; usually it is the rate (per some unit time) of an observed event that is reported and used in the analysis, to account for the potential difference in the observed time (for example between person A and person B).

- Duration or percent spent (of/in a state)--by measuring the length of time the behavior of interest is taking place. Such behavior is commonly referred to as a state (Altmann, 1974).
- Intensity (of a state)—by measuring the degree of the construct's (or behavioral) intensity. This is typically a question of a state's intensity.

Whatever the method you choose to quantify the construct or the behavior of your interest, it will dictate this variable's measuring of scale, and thus, the choices of statistics you will have to test your hypothesis. For example, a researcher can measure bullying by counting the number of times it was observed in a given period, the *ratio scale*, or he/she can measure its intensity, the *ordinal scale* and finally, a researcher can simply categorize it by type: physical vs. verbal, the *nominal scale*. The choice of the measuring scale will have to depend on the research question. Given that one of the research goals in the bullying study by Hawkins, et al. (2001), was to learn how frequently bullying happens in schools, it had to be measured on a ratio scale, ranging between 0 (i.e., no bullying) and some specific observed number of bullying.

Variables

If a hypothesis makes a prediction that one study variable contributes to the occurrence of another variable, the former is the *predictor* and the latter is the *criterion* variable. Suppose that a researcher hypothesizes that the number of books a family has at home predicts the frequency with which preschoolers read at home with their caregivers. Notice that this isn't a causal hypothesis—it only predicts that somehow having more books at home can predict preschoolers' reading habits and, there could be different explanations for that. The number of books in this example is the predictor and the frequency of reading at home is the criterion variable. However, it is important to understand that while an observational study cannot make causal predictions, this doesn't mean that there is no causal connection between the observed events.

It is also possible that a study doesn't test any specific predictions, and that its main goal is to describe, in which case, there is no predictor or a criterion variable. No hypothesis was made in the bullying study, by Hawkins, et al. (2001), about what would predict or would contribute to bullying. The research questions were more of a descriptive nature. The research meant to learn about the frequency of bullying, to find out who/ how often, who would intervene during bullying, and, finally, how effective those interventions were.

Operational Definition and Coding Scheme

Similar to content analysis, observational studies tend to deal with many subjective and somewhat vaguely defined phenomena. And, just like in content analysis, these ***constructs*** have to be ***operationalized*** so that researchers can study them scientifically. Take, for instance, bullying. Although we may think we can all recognize it when we see it, without defining it in very specific

and measurable terms, some forms of bullying may go unrecognized, or some forms of aggression may be mistaken for bullying. As a result, a researcher will not get an accurate measure of the construct. Sometimes, depending on the complexity of the phenomenon and the environment in which it is being observed, a researcher may need to add more specificity, called behavioral categories, and more rules to the operational definition to create a ***coding scheme*** in order to give even more objectivity and standardization. Researchers can borrow an already developed and previously tested operational definition and a coding scheme from other studies, if they provide a reliable measure of the construct of one's own study. If I were to conduct an observational study to further investigate school bullying, I could apply the operational definition and the coding scheme developed by Hawkins and his colleagues (2001). Of course, I would have to appropriately cite the article or the manual where they published the coding scheme to give them their well-deserved credit.

Let's take a look at how Hawkins and colleagues (2001), operationally defined and coded bullying for their study. Bullying was defined as an episode of *aggression*—the intent to inflict pain physically, verbally, or through other implicit means—where the aggressor is more *powerful* physically and/or socially than the victim(s). Notice, that, both constructs, *aggressive episode* and *what is like to be powerful* is given an operational definition, since both are subjective constructs.

An operational definition of a construct doesn't necessarily have to be all-encompassing. It can describe a construct from only one specific point of view as long as its description fits the purpose of the study to reflect the aspect of the construct that is under investigation, and as long as it is objective enough for two or more observers to agree that what they observe is the right construct. If a researcher is only interested in studying the physical forms of bullying, the operational definition needs to provide the objective measures of only physical bullying. The operational definition should be specific and clear enough for multiple observers to all agree on physical bullying when they see it. But is there a way for researchers to know when their operational definitions are objective enough to render reliable measures?

Inter-Rater Reliability

To know how reliable one's operational definition and the obtained measures of the construct can be, researchers will measure the degree of agreement between two independent observers. The simplest index of inter-rater reliability is calculating *percent agreement*, dividing the total number of times observers agreed by the total number of observations, and multiplying it by 100. If observers agreed on what they observed 20 out of 25 observations, the percent agreement between them is 20/25 * 100 = 80%. Eighty percent of agreement can be seen as a large enough number to feel that the obtained measures are reliable. However, the problem with this type of assessment is that it doesn't take into consideration the amount of time the observers can reach agreement purely by chance, and thus, overestimates the extent of agreement. Today, a more acceptable index of

inter-rater reliability is to use *Cohen's kappa* which accounts for the overestimation of agreement due to chance and evaluates the amount of agreement relative to agreement that is expected by chance alone (Bakeman & Gottman, 1997). This statistic can be easily computed using an SPSS program. However, this analysis is only appropriate if your data is nominal. For example, suppose we obtained the following observations of bullying from two observers who independently observed the same events at the same time and recorded the instances of bullying—'0' means no bullying and '1' bullying occurred per event (in this example, ID indicates each event) (see Figure 1).

	ID	Observer1	Observer2
1	1	1	0
2	2	1	1
3	3	0	1
4	4	1	1
5	5	1	0
6	6	1	1
7	7	0	0
8	8	1	1
9	9	0	0
10	10	1	1

Figure 1. *Example of Reliability Data Collected by Two Independent Observers.*

The next step is to determine the inter-rater reliability of the observations by computing Cohen's kappa statistics. Do the following: go to Analyze >> select Descriptive Statistics >> Crosstabs.

Select and move Observer1 to Row(s) and Observer2 to Column(s) (see Figure 2).

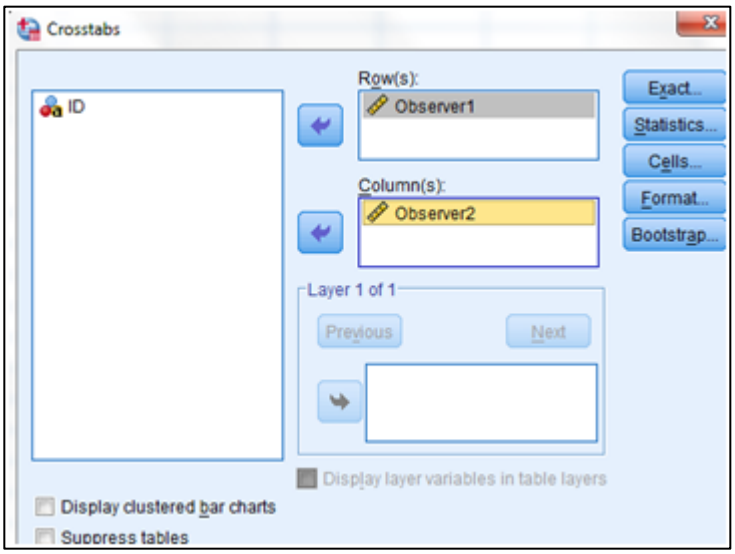


Figure 2. *SPSS Crosstabs.*

Select Statistics option >> Kappa >> continue >> OK

Open the output and scroll down to the table Symmetric Measures (see Table 1).

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.348	.300	1.127	.260
N of Valid Cases		10			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Table 1. *SPSS Output of Kappa Statistic*

The value of kappa is shown to be .348, which is about 35% agreement. The value of kappa can range from -1, which means there is a perfect disagreement between the two observers to +1, which is a perfect agreement; a kappa of 0 indicates only a random agreement between two observers. In social sciences we never expect to have a perfect '1' agreement. An agreement of .60 and .70 is considered an acceptable level. Thus, the kappa value of .34 is very low and indicates that either the operational definition/coding scheme wasn't reliable enough or that one or both observers need more training to obtain more accurate measures and have closer agreement.

Unit of Analysis, Observation, and Coding

When conducting any research, including an observational study, you have to be clear about the focus of your analysis and what or who you will be making conclusions about –you need to know what your ***unit of analysis*** is. If a researcher wants to be able to make conclusions about the types and the frequency of school bullying, the unit of analysis is bullying. On the other hand, if a researcher is more interested in understanding who becomes a bully and who is more likely to be victimized, the units of analysis are the students participating in the study. It is also possible to have more than one unit of analysis. If a researcher is interested in studying the link between bullying and school performance, but would also like to compare instances of bullying and students' school performance across several schools, then there are two units of analysis: the schools and the students.

The ***unit of observation*** is what or who is observed and whose behavior needs to be recorded. Usually, it is the same as the unit of analysis—the people and their behavior—but it certainly doesn't have to be that way. In an observational study the unit of observation has to be based on how the construct that you are trying to capture is operationally defined. For example, if a researcher defines aggression as any act of violence, the unit of observation is an individual committing an act (or several acts) of violence. However, if aggression is defined as an act of violence toward another individual, then the unit of observation must be two individuals where one is committing an act of violence and another is receiving/reacting to it.

The ***unit of coding*** is the portion of the observation that gets a numeric value. In other words, when the behavior being observed is too complex or too long, a researcher will typically break it down into smaller, more manageable units. Each unit is then assigned a number, representing fully or partially the construct of interest. For example, a researcher can set further limits on how long the behavior of a given individual or individuals (i.e., the unit of observation) must be observed--an entire duration of the behavior or just some amount of time. The definitions and other details about what constitutes the unit of analysis, unit of observation, and unit of coding are usually included in the *coding scheme.* The observer is usually trained to understand and follow the coding scheme before he/she begins the observational phase.

To have a better understanding of the role that the units of analysis, observation and coding play in an observational study, let's take a look at what they are in Hawkins, et al. (2001), study about peer interventions in bullying. The unit of analysis in this study was each participant—the children, i.e. the bullies, victims, and interveners, and their behavior, i.e. bullying, being bullied, and peer intervention. The unit of observation was the child bullying another child—so there have to be at least two children based on the operational definition of aggression. Finally, an entire interaction was assessed on whether or not it fit the operational definition of a *bullying episode*. This was the unit of coding. Each bullying episode was coded for the following: the gender of the bully, the victims, and the intervener, the power of the bully (rated on a 5-point scale), the types of peer

interventions (physical, verbal or social), the number of interveners, and whether the victim asked for help.

Can you identify the variables that the coded observations produced and their measurements of scale?

(see answers below)

Appendix: Answers to Practice Exercise

The variables and their respective scales of measurement are:

1. Gender of bully—nominal scale (male and female)
2. Gender of victim—nominal scale (male and female)
3. Gender of intervener—nominal scale (male and female)
4. Power of bully—ordinal or interval (5-point scale)
5. Types of peer interventions –nominal (physical, verbal, and social)

Guided Practice (not graded)

Suppose that you are interested in finding out whether preschoolers exhibit behaviors that are predictive of their later peer interactions in school. Specifically, you want to understand whether they will become a bully, will be at risk of being victimized, or neither one. You hypothesize that a child taking away toys from a peer without that peer's approval is predictive of bullying. So, your first step is to obtain very detailed descriptions of preschoolers' typical social interactions, which may be predictive of their later peer interactions in school, including bullying.

a. So, your first task is to decide on whether you should conduct a naturalistic or a structured observational study to obtain your data.
b. You would like to record the following: the number of times that every preschooler in your study attempted to take away a toy from another preschooler (i.e., aggressive act), the number of times that preschoolers became emotionally distressed regardless of the reason, and the duration of time of the aggressive acts and the emotional distress. What are your variables and their scales of measurement?
c. What is your unit of analysis, unit of observation, and unit of coding?

Answers to Guided Practice Exercise

a. Naturalistic
b. The variables and their respective scales of measurement:
    - Aggressive acts—ratio (can range between 0 and some number)
    - Emotional distress—ratio (can range between 0 and some number)
    - Length of aggressive acts (e.g., average)—ratio (can range between 0 and some number)
    - Length of emotional distress (e.g., average)—ratio (can range between 0 and some number)
c. Preschoolers will be the units of analysis. Two or more preschoolers, and one attempts to take away a toy from another (the unit of observation). The entire aggressive act and the entire episode of an emotional distress are two separate units of coding.

References

Anderson, C. A., & Dill, K. E. (2000). Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. *Journal of Personality and Social Psychology*, 78, 772-790.

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge, England: Cambridge University Press.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119, 593–609.

Hawkins, D. L., Pepler, D. J., & Craig, W. M. (2001). Naturalistic observations of peer interventions in bullying. *Social Development*, 10, 512-527.

Northcote, J., & Livingston, M. (2011). Screening and Identification: Accuracy of self-reported drinking: observational verification of 'Last Occasion' drink estimates of young adults. *Alcohol and Alcoholism, 46*, 6, 709-713.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3, 131–141.

Zeskind, P.S., & Collins, V. (1987). Pitch of infant crying and caregiver responses in a natural setting. *Infant Behavior and Development*, 10, 501-504.

Chapter 6: Experimental Research

Introduction

Experimental design is the quintessential method of a scientific inquiry because it is the ***only method*** that can address the ultimate question of ***causality***. To understand what makes this method so special we will take a closer look at three types of experimental research, within-subjects, between-subjects, and mixed (factorial) design. Each of these experimental methods control and manipulate variables to establish causal relations, and these are the features that give experimental methods advantage over non-experimental research. We will also discuss important differences between the three designs. Specifically, we will go over advantages and disadvantages of using each design as these factors will dictate which one of the three will fit your study the best. Finally, you should know that these experimental designs are only the most basic and common ones researchers use in experimental research. However, all other, more complex and specialized designs, are built on the logic of the ones we will cover in this chapter.

6.1 The Main Features of a True Experiment

What sets the experimental method apart from any other method is the extent of control that researchers have over their study variables. Recall from chapter 3 that in order to establish a cause and effect connection, researchers must be able to (1) manipulate their independent variable and (2) control the effect of all other extraneous variables on the dependent variable. A study is considered an experiment when both requirements are met—the independent variable is being systematically manipulated and all extraneous variables are under researcher's control.

Manipulation of an Independent Variable (Factor)

To establish a cause and effect relationship between two variables, an independent variable (also known as a factor in experimental research) has to be manipulated. Manipulation is nothing more than a systematic change in the level of the independent variable; it can be done quantitatively, by changing its quantity, or qualitatively, changing the kind of factor that is being administered to the subjects. A classic example of a quantitative manipulation is when a researcher administers different *levels* of the *same* drug and observes the effect of each level on the dependent variable. Alternatively, an experimenter may administer different kinds of drugs to the subjects or participants, in which case, the manipulation is qualitative. The drug, in both examples, is an independent variable that is being systematically manipulated. The different levels of an

independent variable are called ***experimental*** or ***treatment*** condition. There must be at least two conditions in an experiment, otherwise, there is no manipulation, either two treatment (experimental) conditions or at least one treatment (experimental) condition and one ***control*** condition. A control condition is a condition where the subjects do not receive any treatment or, instead of treatment, receive a placebo. An experiment can have more than one independent variable or factor, in which case it is called, ***factorial design***.

Controlling Effects of Extraneous Variables

It is not uncommon for many phenomena, which are subjects of psychological investigations, to have multiple causes or to be influenced by various settings, people, etc. So, to establish high internal validity, to be certain that the observed change in the dependent variable was caused by the manipulation of that independent variable and not by something unwanted present in the environment at the time of the experimentation, all extraneous variables must be held unchanged or eliminated from the experiment, if possible. Researchers can accomplish this by (1) randomly assigning subjects or participants to experimental and control groups and (2) by including and measuring or excluding extraneous variables, known to influence the dependent variable.

Random Assignment to Groups

Random assignment is accomplished when the subjects are assigned to different levels of treatment or different group conditions (hereafter I will use *levels* and *conditions* interchangeably as they both refer to the concept of experimental manipulation) by chance alone(not based on certain characteristics or preferences). Random assignment helps eliminate the effect of those extraneous variables that are associated with the subjects' individual characteristics, personality, age, attitudes, demographic/cultural background, and others, by randomly distributing them across groups. Random assignment ensures that all groups, experimental and control, will have an equal mix of all different personal characteristics. In a sense, all groups will have participants of similar characteristics. And if all groups are similar, any observed differences between the groups in the dependent variable can be only attributed to the manipulation of the independent variable, *not* due to differences between the groups in the personal characteristics of the participants.

Including/Excluding Known Extraneous Variables

Certain factors can influence your dependent variable. If a researcher studies a link between aggression and violent video games, gender might be one of the extraneous variables that can independently influence the level of participants' aggressive behavior because we know that, on average, men tend to be more aggressive than women. To eliminate the effect of gender bias, a researcher can study only men or have an equal number of men and women in all groups. The

second technique is called ***matching***, those characteristics that are believed to be influencing the dependent variable, e.g., gender, are matched across the groups. Researchers can also control extraneous variables ***statistically***; however, they would have to measure and include them as additional independent variables in the statistical analysis. Many statistical tests can then analyze the data by statistically removing the effect of extraneous variables. Knowing that certain personality characteristics can affect aggression, a researcher can measure every participant's personality traits and include them in the analysis together with the main independent variable, in this case, video game play.

There are three types of experimental research. The following will qualify the differences and/or similarities in the approaches that each method takes to achieve high internal validity through manipulation and control.

6.2 Within-Subjects Design

When the *same* people or animals are tested across *all* levels of the independent variable, the study is a *within-subjects design*. Below is a real example of a within-subjects experiment, presented to understand its distinctive features.

Example of a Within-Subjects Experiment

Researchers, Domes and colleagues (2007), conducted an experiment to address the following question: Can oxytocin—a chemical, produced in the brain that plays an important role in childbirth, lactation, parenting behavior, and various other social interactions, improve people's understanding of others' thoughts and emotions, the ability also known as mind reading? Although, mind reading may sound like it belongs outside the realm of science, it is actually one of the most widely researched topics in developmental psychology. All healthy people (and even some animals) can understand others' feelings and thoughts, to some extent, without asking them about their thoughts or feelings directly. However, what is also known is that some people are better at mind reading than others. So, what the researchers of the article wanted to know was whether there is a causal connection between the level of oxytocin in the brain and in a person's ability to intuitively understand the minds of other people.

To test this causal question, all participants took part in both the placebo and the treatment condition which were conducted one week apart. In the treatment condition they were given a single dose of oxytocin; in the control condition, a placebo. After the administration of the drug or placebo, all participants were tested by taking a Reading the Mind in the Eyes Test (RMET). This test entailed identifying emotions or thoughts of people depicted on a computer screen. To make

the task as close to mind reading as possible, everything but the depicted people's eyes were purposefully concealed. So the participants had to "read" the thoughts and the emotions of the depicted individuals solely based on the expressiveness of people's eyes.

The variables of this experiment were the following:

The *independent* variable was the chemical, oxytocin, with two levels—experimental and control conditions.

- Experimental condition (administration of oxytocin)
- Control condition (administration of a placebo)

The *dependent* variable was the scores on the "mind reading" test (RMET). The same participants underwent the experimental *and* the control conditions, which makes it a *within-subjects design*.

Strengths of a Within-Subjects Experiment

A within-subjects experiment requires fewer participants than a between-subjects design because the same people or animals proceed through all levels of the experiment. In the experiment we just reviewed, the same thirty, male students participated in the experimental and control conditions. So, the study required the participation of only thirty individuals in total.

Another important advantage of a within-subjects design is the fact that researchers do not have to control for any individual differences that may be present between groups because the *same* people, or animals, comprise all of the groups. Thus, there are no individual between-group differences. Recall, that these differences can potentially confound the results of a study if they are not properly controlled which will decrease internal validity. People may differ in their ability to understand the emotions and thoughts of others as a result of many different factors. So, to be confident that the improvement in their "mind reading" is attributed to the administration of oxytocin, and not due to other factors such as differences in how individuals were raised or their innate ability to understand other's emotions, the same participants were tested in both conditions, effectively eliminating any other alternative explanations for the observed change in the dependent variable.

Weaknesses

A within-subject design is not without some problems. Testing the same people over and over can result in *carryover, maturation, or history effects.*

Carryover Effects

The carryover effect happens when the effect of one condition carries over or influences the performance of the subjects in the next condition. Some of the most common reasons for the carryover effect are *order* and *timing* between conditions, *practice* and *fatigue*.

Order and timing are major contributing factors to testing bias. Suppose the subjects are administered drug A in the first condition, and drug B in the next condition. It is likely that the effect of drug A will continue exerting its effect on the subjects in the next condition or interact with drug B if the conditions are not properly ordered or if the timing between the conditions is not sufficient for the effect of drug A to dissipate. In another example, imagine that researchers are testing the effect of violent video games on game players' aggression. First, the subjects play a non-violent video game, the control condition, and, then, they play a violent video game, the experimental condition. Any increase in the subjects' physiological arousal of aggression can be attributed to at least two factors: the effect of playing a violent video game, the independent variable, and the effect of playing two video games for an extended period of time, described as the carryover effect. The problem is that we would not be able to separate the two effects and the results would be most definitely confounded.

Practice and Fatigue are also potential threats to internal validity because they influence participants' performance and confound the true effect of the independent variable. For example, the *practice effect* would have been present if the participants in Domes et al. (2006)'s experiment had been tested in the experimental condition right after the placebo condition. Their performance would have been improved even without the administration of oxytocin, simply due to taking the same test twice. *Fatigue* works in the opposite direction because it decreases participant performance. This usually happens when subjects are being tested over an extended period of time, or when the test requires high concentration or is monotonous.

Maturation

Maturation occurs when a naturally-resulting internal process, such as aging or learning, and not manipulation of an independent variable, causes a change in the dependent variable. This presents another threat to internal validity when it comes to a within-subjects experiment. This is particularly problematic when the study is longitudinal, continuing over an extended period of time. Assume we want to test the effectiveness of an intervention program, designed to improve preschoolers' attention. If the program continues over 6 months, any improvements in the preschoolers' attentional focus could be simply attributed to the maturation of their brain rather than to the effectiveness of the program.

History Effect

Finally, an effect of some external event, other than manipulation of the independent variable, that causes the change in the dependent variable, is known as *history effect.* Similar to *maturation*, it poses a more significant threat to internal validity if it is a longitudinal study and more unexpected and uncontrollable events can happen in the course of the study. Suppose a clinical psychologist decides to conduct a within-subjects experiment to test the effectiveness of a therapy that treats anxiety. To ensure that the therapy has a lasting effect, the psychologist waits one month after the patients received the treatment to conduct a post-treatment to measure their anxiety level. He wants to see whether the treatment continued to be effective. Now imagine that during the waiting period something dramatic has happened, like a natural disaster, and the patients witnessed it. It is quite possible that the patients' anxiety would spike again, even if the treatment did work. This would undermine the entire experiment because the effect of the therapy would be confounded by the effect of the natural disaster.

Ways to control threats to internal validity of a within-subjects design.

Some of the afore mentioned confounding factors can and should be controlled when conducting a within-subjects experiment. Practice and fatigue can be combated in many cases by properly separating the conditions. For example, in the Domes et al. (2006)'s experiment, the experimental and the control conditions were completed within one week to avoid potential practice or fatigue effect. Maturation and history can be effectively controlled by adding a control condition. Thus, for example, the control group in the intervention program example would be influenced by maturation as well. If differences in the outcomes are still found, they could be attributed to the intervention program only. Similarly, both the control and the experimental conditions in the therapy example would witness the same, historic event. However, if any differences between those groups were still found, they would be due to the effectiveness of the therapy only.

Counterbalancing the order of the conditions is another way to control for the carryover effects. Although all participants will go through all levels of the independent variable, they will do so in a different order. Suppose a researcher wants to compare the effectiveness of two different treatments: A vs. B. To control for the factors we have just discussed, the researcher will *counterbalance* the order of the conditions: two participants will undergo treatment A first and the other two will start with treatment B (see Table 1).

Table 1. Full Counterbalancing

	ORDER OF CONDITIONS	
	First	Second
Participant 1	A	B
Participant 2	A	B
Participant 3	B	A
Participant 4	B	A

The logic behind counterbalancing is the following: Any systematic influence of the AB order will be counter-balanced by the BA effect. In other words, they will cancel each other out.

Fully counterbalancing orders of conditions, i.e., including all possible orders in the study, can be easily accomplished when there are only two conditions (e.g., A and B). But counterbalancing three or more conditions gets exponentially complicated. To fully counterbalance three conditions—A, B, and C—six order conditions will have to be included (i.e., permutation, 3!=3x2x1): ABC, ACB, BCA, BAC, CAB, and CBA. A study with 4 levels of an independent variable will have 24 different order permutations (4! = 4x3x2x1). So to make the study more manageable, ***partial counterbalancing*** can be a good alternative. One of the common partial counterbalancing techniques is a, ***Latin square***. A Latin Square is where each condition appears in every order, but only once (i.e., first, second, third, etc.). Thus, an experiment with A, B, and C conditions can be partially counterbalanced as, ABC, BCA, and CAB. All three conditions have appeared once as first, second, and third (see Table 2).

Table 2. Partial Counterbalancing: Latin Square

	ORDER OF CONDITIONS		
	First	Second	Third
Participant 1	A	B	C
Participant 2	A	B	C
Participant 3	B	C	A
Participant 4	B	C	A
Participant 5	C	A	B
Participant 6	C	A	B

Finally, if none of the strategies can prevent the confounding that we have discussed in this section, or if the strategies cannot be implemented due to methodological, ethical or other reasons, a researcher will have to conduct a *between-subjects* experiment.

6.3 Between-Subjects Design

An experiment is said to have a between-subject design if different subjects or participants receive different levels of an experimental manipulation. If an experiment calls for three experimental, and one control condition, there will be four groups: three experimental and one control group. Each group of subjects or participants (randomly assigned to one of the four groups) will receive different levels of the treatment or no treatment if it is a control group.

Example of a between-subjects experiment.

Elmore and Oyserman (2012), tested an *identity-motivation theory*, according to which, people are motivated to act in accordance with what they consider congruent to their own identity behavior. This applies to gender-identity as well. Furthermore, according to this theory, people pick up clues on what their identity-congruent behavior is supposed to resemble from their environment. Thus, boys and girls will learn behaviors appropriate to their gender partially from the information they receive in their surroundings. The researchers of this article hypothesized that school-focused behavior (behavior that promotes academic success) will be affected by whether or not the contextual clues link their gender with school success. The dependent and independent variables in this experiment are as follows:

There are two *dependent variables* (outcomes) in this study:

- School-focused behavior— a description of oneself that includes academic themes, such as expecting or wanting to get good grades, being or trying to behave well in school, or a mentioning of any behavior or thoughts that would contribute to school success.
- Future success expectations—a narrative that includes future success expectations.
- Math task—a number of attempts the student made to solve a math problem.

All dependent variables were expected to be affected by the contextual clues, a.k.a. the independent variable.

The *independent variable* was visual information students received about what gender congruent or incongruent behavior was supposed to be; the following levels were manipulated:

- Experimental condition 1: male and female students were randomly assigned to receive a graph displaying men earning more money than women (*boy- congruent*success clue

since men, on average, earn more than women). The hypothesis was that this visual clue would motivate boys to adopt school-focused behavior and to make more attempts to solve the math problem.

- Experimental condition 2: students were randomly assigned to receive a graph displaying more girls graduating from high school than boys (*girl-congruent*success clue since women, on average, are more likely to graduate from high school). The hypothysis was that this visual clue would motivate girls to adopt school-focused behavior and to make more attempts to solve the math problem.
- Control condition 1: students were randomly assigned to receive a graph displaying the typical household income in Michigan with no gender-related information.
- Control condition 2: students were randomly assigned to receive a graph displaying the high school graduation rate in Michigan again, with no gender-related information.

*To support the identity-motivation theory, *both* male and female students in gender congruent conditions should have higher scores on the dependent variables than those male and female students who were in gender-incongruent conditions.

Such group differences, in fact, were found: *Gender-congruent success clues influenced both girls' and boys' school-related behavior and expectations. Students of both genders described themselves as having or imagined themselves as possessing more school-focused identities, made more attempts to solve the math problem, and had higher future expectations for themselves.*

Advantages

The confounding factors which threaten internal validity of a within-subjects design, practice, fatigue, maturation, history, and carryover effects, are usually absent from a between-subjects experiment although, they can be present in some between-subjects experiments. So, this experimental design is a good alternative to a within-subjects experiment when no other strategies can work to eliminate the confounding effects that are associated with repeatedly testing the same subjects over the course of an experiment.

Disadvantages

Although random assignment of participants or subjects to the groups of an experiment theoretically should eliminate any confounding variables, such as individual differences across groups that are not due to manipulation of an independent variable, the fact remains that some differences may still remain and can influence the outcome of the study. Any systematic and uncontrolled differences between groups that are not due to manipulation of an independent variable can undermine internal validity of an entire experiment.

Finally, this design usually requires more participants (or subjects) since different people, or animals, will be tested in every condition.

Final Notes on When to Choose a Within- or a Between-Subjects Design

Start with the goal of your research—what is the question that you are trying to address? If you are interested in the effect of practice or of repeated exposure to a certain stimulus, then a within-subjects design is the method of choice. However, if you are testing the effectiveness of different learning strategies, practice effect is no longer the desirable strategy and will have to be controlled either by counterbalancing the orders of the conditions or by spacing them out properly over time.

Related to this is the issue of *external validity.* The design with which you choose to study your phenomenon should resemble the context in which it occurs in real life. If one wants to study how media credibility affects their likelihood of persuading an audience (trustworthy media vs. non-trustworthy media source), a within-subjects design reflects the reality of media influences on its audience more accurately than a between-subject study *The same* people tend to be exposed to different (credible and non-credible) media sources (e.g., television, radio, Internet). See Greenwald, 1976, for more details. However, if you believe that the design you chose will significantly compromise the internal validity of your study, you may need to select the between-subjects design. In many cases, there will always be a trade-off between internal and external validity and you will have to choose a method that can minimize the number of potential confounding variables while still maintaining a reasonably good external validity.

It also helps to review the empirical literature on the same topic to find out what is the most accepted design to study the same or similar phenomenon. These studies will generally inform you about any methodological issues associated with the methods that have been already applied. They may even make some useful recommendations that you can implement in your own study.

The number of subjects that are available for participation or can logistically be recruited is another factor to be considered. As previously discussed, a within-subjects experiment will require fewer people or animals than a between-subjects experiment since the same subjects will go through all levels of a study.

Finally, a within-subjects design is the best choice when individual characteristics exert a significant effect on the dependent variable of the experiment (Bordens & Abbott, 2011). Since they can be difficult or, in some cases, impossible to control, having the same individuals go through all the conditions of an experiment will resolve this problem automatically.

6.4 Factorial Design

Factorial

So far the examples in the previous section involved only one independent variable with two or more levels. But more often than not, studies will include more than one independent variable in order to more accurately explain the phenomenon under investigation or to address several questions in one study. Any experimental design (a between- or a within-subjects design) with more than one independent variable is called a ***factorial design***.

Here is an example of a between-subjects *factorial* experiment:

Ivory and Kalyanaraman (2007), set out to test several related hypotheses that can be summarized as follows: Newer and violent video games exert stronger influence on the players than the older, violent, or nonviolent video games. Thus, in addition to testing the effect of *violence* on game players' emotions and behavior, Ivory and Kalyanaraman (2007), included a second factor,*technological advancement*. This was included to find out if both or just a single independent variable would significantly affect the participants. The design was also a between-subjects experiment. The variables of this experiment are as follows:

The *independent variables* of the experiment and their levels are:

1. Video game violence with two levels
    1. Playing violent video games
    2. Playing non-violent video games
2. Video game technological advancement with two levels
    1. Newer video games
    2. Older video games

The *dependent variables* of the experiment are:

1. Presence (perceptual experience of being in the game)
2. Involvement (being engaged in the game)
3. Physical Arousal
4. Aggressive thoughts
5. Aggressive feelings

This is a 2x2 (two by two) factorial design, where the first number '2' refers to the levels of the *first* factor (violence with two levels) and the second number '2' identifies the levels of the

second factor (technological advancement with two levels). For convenience, this design can also be represented as a tabulation. (see Table 3):

Table 3. Conditions of A 2x2 Factorial Design

	Factor 2: Violence	
Factor 1: Advancement	**(Level 1) Violent**	**(Level 2) Non-Violent**
(Level 1) Newer	Playing *newer violent* video games	Playing *newer non-violent* video games
(Level 2) Older	Playing *older violent* video games	Playing *older non-violent* video games

A factorial experiment of 3x2x4 would mean that it has three independent variables; the first independent variable has 3 levels, the second independent variable has two levels and the third has 4 levels.

In a factorial experiment, there are two types of effects (or outcomes or results) that researches will look for: *main* and *interaction* effects.

Interpreting Results of a Factorial Experiment:

Main Effects

This is the effect that each independent variable exerts on the dependent variable regardless of the effect of other independent variables in a study. Let's go back to Ivory and Kalyanaraman (2007), were five main effect hypotheses were to be tested (two with violence and three with advancement).

Main Effects of Violence:

1. Violent video games will cause higher levels of physiological arousal in the participants than will nonviolent games.

2. Violent video games will cause higher levels of aggressive thoughts than will nonviolent games.

In other words, both predictions state that *violence* will *independently* produce a change in the physical arousal and the aggressive cognition, regardless of the games' technological advancement.

Main Effects of Advancement:

1. Newer video games will cause higher levels of presence in game players.
2. Newer video games will cause players to feel more involved with the game than will older games.
3. Newer video games will cause higher levels of physiological arousal in game players.

In other words, the researchers hypothesized that *advancement* would exert a main effect on three dependent variables—*presence, involvement,* and *physical arousal*.

The experiment, in fact, confirmed the main effects of advancement only.

I will use Ivory and Kalyanaraman (2007) study to simulate a hypothetical data to show how a main effect can be detected.

First, let us specify the conditions of this experiment and create a table where we could input hypothetical means from each condition. This is a 2x2 factorial design with 4 conditions: Violent and New, Violent and Old, Non-Violent and New, and Non-Violent and Old. We can also assume that every participant was randomly assigned to only one of the four conditions, that is, the between-subjects design. Suppose that we have the following scores (i.e., averages) from each groups (see Table 4).

Table 4. Mean Physical Arousal Scores by Violence and Advancement when Only Advancement has a Main Effect

	Violence		Marginal Means
Advancement	(Level 1) Violent	(Level 2) Non-Violent	
(Level 1) Newer	35 (Group 1)	35 (Group 2)	35
(Level 2) Older	10 (Group 3)	10 (Group 4)	10
Marginal Means	*22.5*	*22.5*	

The *marginal means* are the means that came from one level of one independent variable averaged across the levels of the other independent variable. Highlighted in red are the marginal means that are significantly different, statistically, from each other. These means represent a main effect of the corresponding independent variable. Thus, 35 and 10 are the marginal means corresponding to the scores of the players' physical arousal due to manipulation of advancement only. This chart illustrates that the newer games (M=35) induce more physical arousal in the players then the older games (M=10). *Keep in mind that we would still need to perform a statistical analysis to confirm what appears to be a significant difference between the marginal means of advancement. Statistical analyses and their computation will be discussed elsewhere.*

Violence, on the other hand, does not have a main effect on physical arousal in this example since both violent and non-violent conditions have the same mean of 22.5.

If we plot the means, the main effect of Advancement and the absence of the main effect of Violence will look like the following (see Figure 1). The two levels of Advancement are marked as blue and red lines, while the violent and the non-violent conditions are represented by the flatness of the lines.

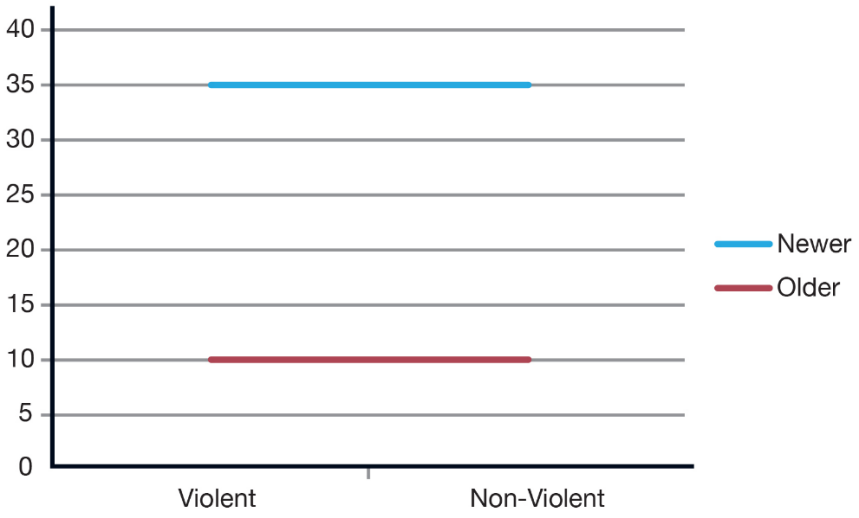


Figure 1. *A Main Effect of Advancement but not of Violence.*

Let's consider a reversed scenario, when violence, and not advancement, has a main effect on physical arousal (see Table 5 and Figure 2). The players in the violent condition have higher physical arousal than in a non-violent condition. Notice that the graph of a main effect of violence looks different from the one in Figure 1, i.e., the line is diagonal representing the violent and non-violent means (M=35 and M=10, respectively).

Table 5. Mean Physical Arousal Scores by Violence and Advancement when Only Violence has a Main Effect

	Violence		Marginal Means
Advancement	(Level 1) Violent	(Level 2) Non-Violent	
(Level 1) Newer	35 (Group 1)	10 (Group 2)	22.5
(Level 2) Older	35 (Group 3)	10 (Group 4)	22.5
Marginal Means	*35*	*10*	

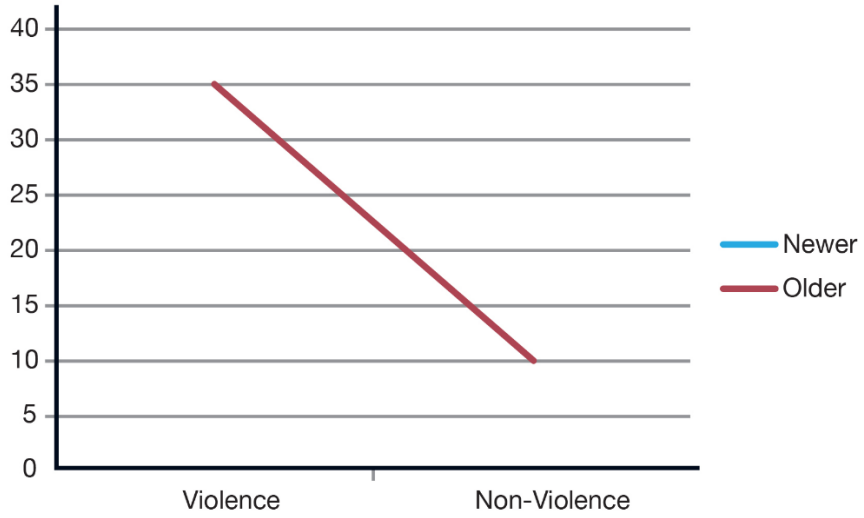


Figure 2. *A Main Effect of Violence but not of Advancement.*

Now, let's suppose that both Violence and Advancement have main effects on physical arousal (see Table 6). Notice that the marginal means of both *advancement* and *violence* are different at both levels. The way to interpret this is to say that, both violence in the video games and the technological advancement of the video games increase, *independently*, players' physical arousal.

Table 6. Mean Physical Arousal Scores by Violence and Advancement when both Factors have Main Effects

	Violence		**Marginal Means (Main Effect of Advancement)**
Advancement	(Level 1) Violent	(Level 2) Non-violent	
(Level 1) Newer	35 (Group1)	30 (Group 2)	*32.5*
(Level 2) Older	10 (Group 3)	5 (Group 4)	*7.5*
Marginal Means (Main Effect of Violence)	*22.5*	*17.5*	

The graphic representation of the main effects for both factors will be as following (see Figure 3). Notice that the two lines are parallel and diagonal.

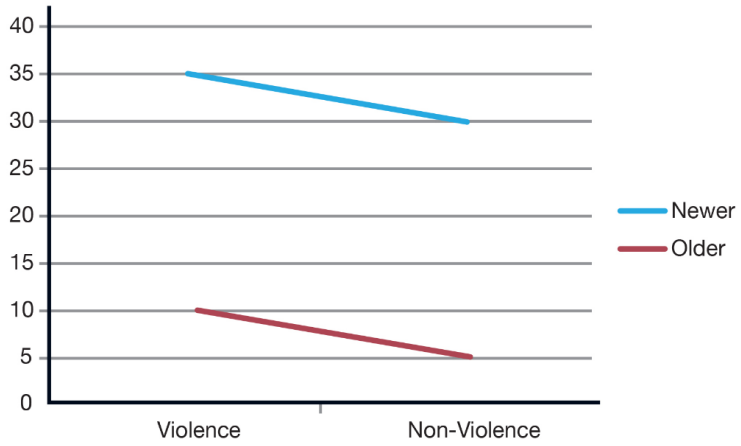


Figure 3. *Main effects of Advancement and Violence.*

Interaction Effects

In a factorial design, independent variables can also interact producing an *interaction effect*. This means that the effect of one independent variable is contingent upon levels of another independent variable. The effect on the players' physical arousal, of playing a violent video game, may be different when the game is newer or older. Similarly, a non-violent video game may affect the players differently if they play a newer or older, non-violent game. Ivory and Kalyanaraman (2007), tested whether the level of physical arousal would be greater in *violent* and *newer* video game conditions than in *non-violent*, newer or older, video game conditions. Although, this hypothesis was not supported, hypothetical data will be used to show what the means and the graph would look like when *interaction* is present in the data.

Suppose that we have the following means (see Table 7).

Table 7. Mean Physical Arousal Scores by Violence and Advancement that Show Main and Interaction Effects

	Violence		Marginal Means (Main Effect of Advancement)
Advancement	**(Level 1) Violent**	**(Level 2) Non-Violent**	
(Level 1) Newer	50 (Group 1)	32 (Group 2)	41
(Level 2) Older	15 (Group 3)	40 (Group 4)	27.5
Marginal Means (Main Effect of Violence)	*32.5*	*36*	

First, one should notice that the marginal means of *violence* and *advancement* are different. We will assume that the differences are also statistically significant. Physical arousal is higher in newer than in the older conditions. It is also higher in the non-violent rather than in the violent condition, which is probably the opposite of what one would predict. The differences in the marginal means of both factors suggest that violence and technological advancement have main effects on physical arousal. However, the examination of the cell means (the *means* of the four conditions) reveals an even more interesting pattern. Physical arousal is highest in the group where the players played newer *and* violent video games; it is the lowest when they played older *and* violent video games.

If we plot the means of the conditions, the lines will be crossed, indicating that the effects of the two variables interact (see Figure 4).

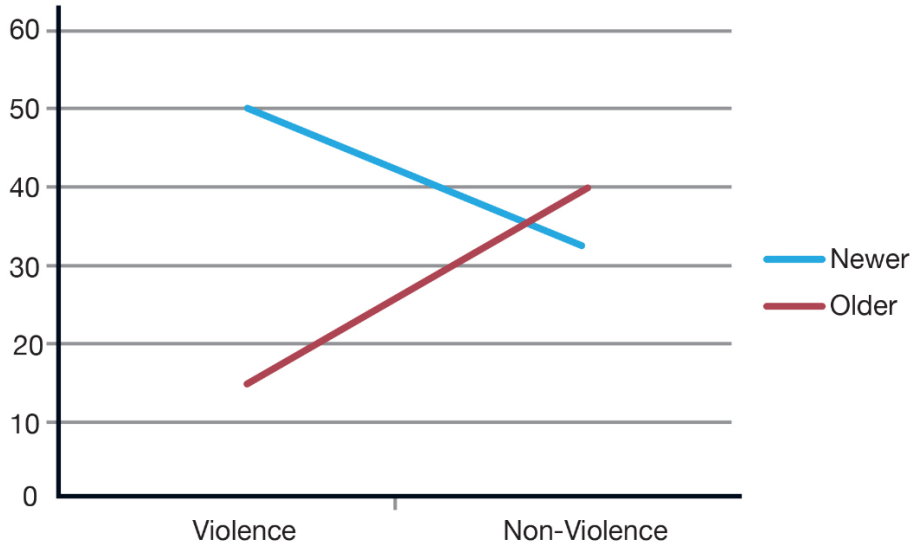


Figure 4. Main and Interaction Effects of Advancement *and* Violence.

Finally, it is possible to have significant interaction effect without a main effect. Consider the following data in Table 8.

Table 8. Mean Physical Arousal Scores by Violence and Advancement that show an Interaction Effect only

	Violence		Marginal Means (Main Effect of Advancement)
Advancement	**(Level 1) Violent**	**(Level 2) Non-Violent**	
(Level 1) Newer	50 (Group 1)	10 (Group 2)	30
(Level 2) Older	10 (Group 3)	50 (Group 4)	30
Marginal Means (Main Effect of Violence)	*30*	*30*	

The marginal means of both factors are the same across their two levels. This indicates that there are no main effects. Physical arousal did not differ between the two levels. Both newer and older conditions have a mean of 30. Similarly, the mean of physical arousal was 30 in both violent and non-violent conditions. However, there is *interaction*: When the players played violent *and* newer or non-violent *and* older games, their physical arousal was increased. Nevertheless, when they played violent *and* older or non-violent *and* newer games their physical arousal was reduced. Without testing for interaction, we would have erroneously concluded that the content of the games, or their technological advancements, had no effect on the physical arousal of the players, when, in fact, they do by interacting with each other. The interaction between the two factors is also evident in the graph where the lines are non-parallel and crossed.

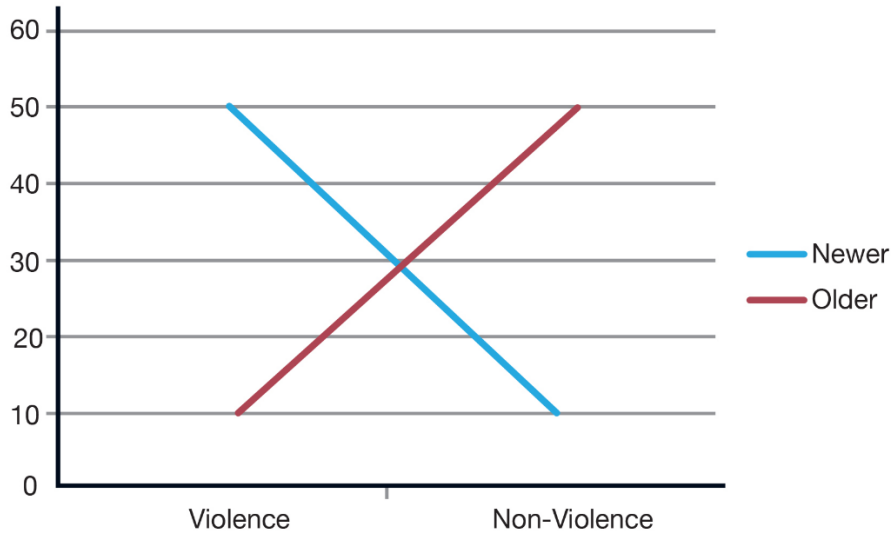


Figure 5. *Interaction Effect of Advancement and Violence but No Main Effects*

Final Notes on Graphically Recognizing Main and Interaction Effects

When you have a *main effect for one of the factors*, the lines will be parallel and separate (see Figure 1); or one line will be diagonal (not flat) (Figure 2). Main effects for both factors are evident when both lines are parallel, separate, and diagonal (see Figure 3). An *interaction effect* is present when the lines are crossed or non-parallel (see Figures 4 and 5).

Mixed Designs

An experiment that combines within- and between-subjects elements is a *mixed design* with at least one within, and one between, subjects variable. There are several reasons for conducting a mixed design study. The following illustrates commonly occurred scenarios when a mixed design is implemented.

- When participants have been naturally self-assigned to one of the levels of the independent variable, this ***quasi-independent*** variable is, by default, a between-subjects factor (e.g., age, gender, disease) because each participant can only undergo one of the levels (e.g., male or female condition only).
    - Example: Let's recall a study by Domes et al. (2006)—a within-subjects experiment which tested a causal link between oxytocin and mind reading. Suppose the researchers added a second factor, gender, to find out if a combination of being a female and having an elevated level of oxytocin improves one mind reading significantly (i.e., additive effect). It is now a 2x2 mixed factorial design with gender as a between- and oxytocin as a within-subject variable (see also Table 9).

Table 9. An Example of A 2x2 Mixed Factorial Design with Oxytocin and Gender

Oxytocin (Within-Subjects Factor)	Gender (Between-Subjects Factor)
(Level 1) Oxytocin	(Level 1) Men
	(Level 2) Women
(Level 2) Placebo	(Level 1) Men
	(Level 2) Women

- In a within-subjects design, counterbalancing can be treated as a second between-subjects factor. To return to the hypothetical example with two treatments: A and B discussed in the section on counterbalancing, the two counterbalanced order conditions (AB and BA) can be treated as two levels of a between-subjects factor since every participant could only take the two treatments in one particular order, either as AB or BA (see Table 10). By including the counterbalanced order conditions as a factor of the analysis, researchers can control and study them.

Table 10. An Example of a 2x2 Mixed Factorial Design with A and B Treatments and Counterbalanced Order Conditions

Treatment (Within-Subjects Factor)	Counterbalanced Orders (Between-Subjects Factor)
(Level 1) Treatment A	(Level 1) AB order
	(Level 2) BA order
(Level 2) Treatment B	(Level 1) AB order
	(Level 2) BA order

- *Pretest-posttest design* is another example of a mixed design, wherein different treatment conditions are the between-subjects factor and, the pretest and posttest is a within-subjects factor. Let's use treatment A and B from the previous example here as well. But this time half of the participants will undergo only treatment A and the other half will take treatment B. Treatment B could be a placebo, a control condition, or, conversely, a different kind of treatment. To be sure that the treatment of interest, treatment A, did work, all participants will be tested before, and after, the treatment and then their scores will be compared (see Table 11).

Table 11. An Example of a 2x2 Mixed Factorial Design with A and B Treatments and Pretest-Posttest Conditions

Oxytocin (Within-Subjects Factor)	Gender (Between-Subjects Factor)
(Level 1) Oxytocin	(Level 1) Men
	(Level 2) Women
(Level 2) Placebo	(Level 1) Men
	(Level 2) Women

References

Bordens, K. S., & Abbott, B. B. (2011). *Research design and methods: A process approach*(8th ed.). New York, NY: McGraw Hill.

Domes, G., Heinrichs, M., Michel, A., Berger, C., & Herpertz, S. C. (2007). Oxytocin improves "mind reading" in humans. *Biological Psychiatry*, 61, 731-733.

Elmore, K. C., & Oyserman, D. (2012). If 'we' can succeed, 'I' can too: Identity-based motivation and gender in the classroom. *Contemporary Educational Psychology*, 37, 176-185.

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin*, 83, 314-320.

Ivory, J. D., & Kalyanaraman, S. (2007). The effects of technological advancement and violent content in video games on players' feelings of presence, involvement, physiological arousal, and aggression. *Journal of Communication*, 57, 532-555.

Chapter 7: Quasi-Experimental Design

Introduction

Sometimes a study is only, *quasi*-experimental or, *partially* experimental, because one or two key features of a true experiment are missing; either the participants are *not randomly* assigned, and/or no control or comparison group is used. Both of these requirements must be met in order to observe a pure, un-confounded effect of an independent variable on a dependent variable. However, it is still the second strongest design in terms of its internal validity after the experimental method. This is because internal validity in a quasi-experimental study is still higher than in any other non-experimental research. In addition, some quasi-experiments, such as natural experiments, have a higher level of external validity than an experimental design because they make use of some *naturally* occurring phenomenon to serve the role of independent variable.

As you will see later in this chapter, the most common reasons why quasi-experimental studies do not follow random assignment protocol or, do not include an equivalent control group are:

- Convenience—No time or resources to recruit a control group or to implement random assignment.
- Unethical—It would be unethical to subject randomly assigned people to the emotionally or physically damaging effects of some independent variables.
- Impossible— Some phenomena, like gender or age, cannot be manipulated, thus, participants cannot be randomly assigned to different levels of these quasi-independent variables.

Regardless of the problems associated with quasi-experimental research, studies of this nature are very important and useful because they allow researchers to study questions that otherwise would not be possible or ethical to perform. Furthermore, they are also quite common in policy and program evaluations.

7.1 Typical Quasi-Experimental Designs

One-Group Posttest-Only Design

Recall a pre-test/post-test experimental design where researchers test at least one experimental group, treatment group A, and one comparison group, comparison group B, on the same dependent variable before and after the treatment where: treatment A for subjects in group A and either no treatment, a placebo, or some other treatment for B. (see Figure 1).

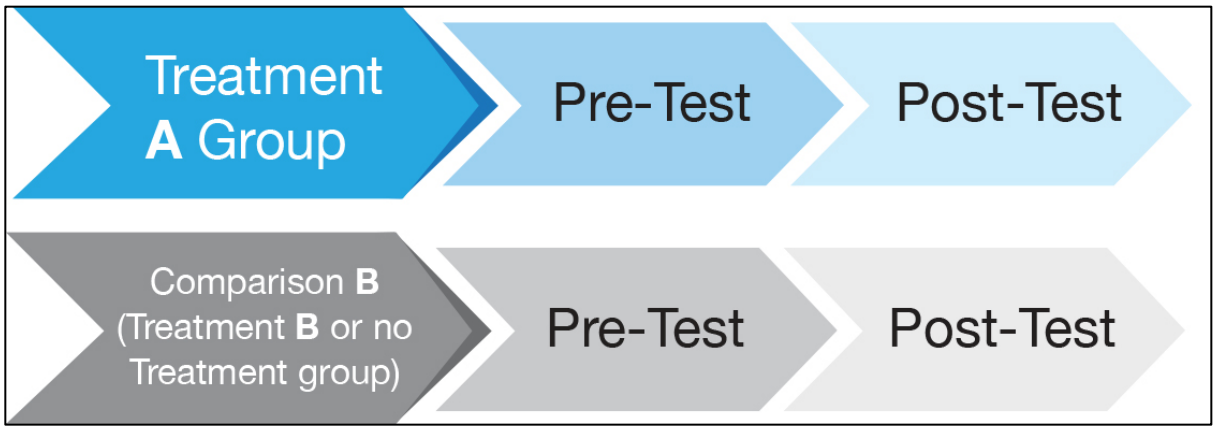


Figure 1. *An Example of a 2x2 Mixed Factorial Design with A and B Treatments and Pretest-Posttest Conditions.*

However, researchers may not always be able to have a comparison group or even a chance to test the subjects before the treatment. For example, if an instructor wants to measure the effectiveness of her lecture, she may simply administer a test to measure her students' knowledge of the lecture after the fact. Since no measures were taken *before* the lecture and because no comparison (or a control) group was used to control for confounding factors, it is a one-group posttest-only quasi-experimental design (see Figure 2).



Figure 2. *An Example of a One-Group Posttest-Only Design.*

Non-Equivalent Control Group Designs

Now, suppose the same instructor wants to increase the internal validity of her study and be able to draw more certain conclusions. So, she decides to compare the results of the same test that examines the *same* course material but that was administered to a different class of students taught by a different instructor. This becomes a comparison group. It is similar to a control group because, it is used to compare the results of the experimental condition. However, it is also different because unlike the control group, where subjects do not receive any treatment or only receive a placebo, the comparison group receives some alternative treatment. This is an example of a non-equivalent control group design, because the subjects are not randomly assigned to the control condition

Likewise, the experimental group in this example is not *randomly* assigned, the instructor used her class, and unless, she only cares about the effect of her lecture on *her* class, there is a problem with the generalization of her results to other students and instructors.

As the above examples might suggest, a non-equivalent control is often implemented out of convenience, i.e., subjects are geographically close to the location of the study, already enrolled in a class or a program, etc.. It is also one of the reasons why many program evaluation studies have a *non-equivalent control design*. For example, people who complete the program (e.g., an intervention or a therapy) constitute an experimental condition and, people who either choose not to participate in the program or who choose to try a different program, are treated as a comparison group. The key reason why such a comparison group is a non-equivalent control is because people were not *randomly* assigned to the experimental and comparison groups.

The subjects in this instance were self-selected, they chose not to take part in the program or to take part in a different program, and that can potentially lead to assignment bias. *Assignment bias* exists where any systematic individual differences between the groups are not due to manipulation of the independent variable. For example, people who choose not to participate in a study may differ from their counterparts in personality, sex, demographic characteristics, severity of their problem, etc. These differences are potential confounding factors because they offer alternative explanations to observed differences in the dependent variable. Thus, this type of research does not permit causal inferences like a true experiment would.

Below are a couple of non-equivalent control group designs:

By adding a non-equivalent control group to a one-group posttest-only design, the study becomes a ***non-equivalent control group posttest design*** (see Figure 4).

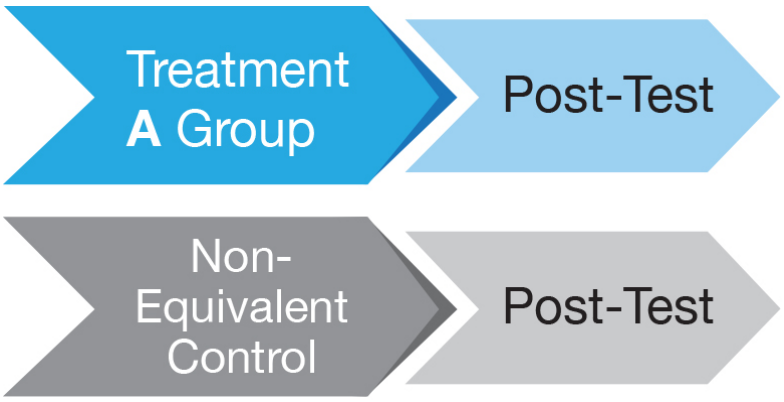


Figure 4. *An Example of a Non-Equivalent Control Group Posttest Design.*

Similarly, if a one-group pretest-posttest study is added, a non-equivalent control group, the study is a ***non-equivalent control group pretest-posttest design*** (see Figure 5).

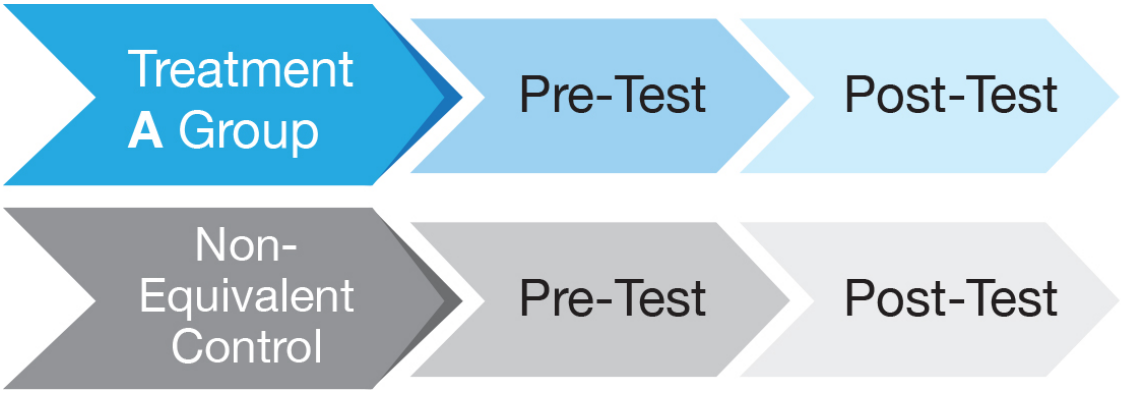


Figure 5. *An Example of a Non-Equivalent Control Pretest-Posttest Design.*

There is a key distinction between a Non-Equivalent Control Group Designs and a True Experiment. Unlike in a true experiment, participants in a non-equivalent control group are not randomly assigned. They are, however, chosen to match as many characteristics of the participants in the experimental condition as possible, to minimize any confounding variables that can be found between groups.

Example of a non-equivalent control group design.

At the beginning of this chapter, I mentioned that sometimes random assignment is not possible for ethical reasons. During a study by Schneider-Rosen and Cicchetti (1991), researchers dealt with just such an independent variable. Specifically, Schneider-Rosen and Cicchetti were interested in examining the effect of maltreatment on toddlers' self-knowledge, or, their awareness of themselves as distinct and separate from other individuals. Their hypothesis was that this social-emotional milestone, which typically develops during the second year of life, would show delay in children who have been maltreated.

The variables of this quasi-experimental study were:

- The independent variable is *maltreatment*. Normally, participants would be randomly assigned to at least two conditions —maltreated and non-maltreated conditions—and subjected to the effect of the independent variable. Since this would be unethical, immoral, and illegal to do, Schneider-Rosen and Cicchetti recruited children who already had some history of maltreatment. Thus, the levels of the independent variable were:

    - Maltreatment during the first two year of the child's life.
    - No maltreatment during the first two year of the child's life.
- The dependent variable is self-knowledge. It was measured using "mirror self-image" paradigm (the details of this technique are described in the article).

Controlling for the Confounding caused by lack of Randomization.

The study by Schneider-Rosen and Cicchetti (1991), is a good example of why self-assignment (e.g., prior history of maltreatment) can lead to *assignment bias* and confounding. The majority of maltreated children who were recruited for the study came from low socio-economic backgrounds (low SES). This demographic is known to influence children's development in many ways, and could also affect the development of self-knowledge. Since the purpose of the study was to understand the effects of maltreatment, low SES was a confounding factor that had to be controlled. Rosen and Cicchetti used a matching technique to do just that. These researchers recruited children for the control condition (i.e., non-equivalent control group) who had no prior history of maltreatment but, of the same age and socio-economic status as the children in the experimental condition. This strategy helped the researchers compensate for the lack of randomization and draw meaningful conclusions about the differences between the two groups in terms of self-knowledge.

7.2 Natural Experiments

Studies that exploit natural events to serve the role of an independent variable are called *natural experiments.* They are quasi-experimental because the subjects are not randomly assigned by the investigator to the experimental groups (e.g., the subjects who experienced the event) and the control groups (e.g., the subjects who did not experience the event), and also because the control group is usually non-equivalent. Natural experiments are commonly used in economics and program evaluations but are becoming more popular in psychology because they can complement both experimental and non-experimental research in many ways. For example, when manipulation of an independent variable is impossible, using some naturally occurred event can be an alternative option. This is similar to using a quasi-independent variable like recruiting subjects who have been naturally self-assigned to one of the conditions such as the study of maltreated children by Schneider-Rosen and Cicchetti (1991). The following examines one such study that used a naturally occurred event to examine the link between poverty and mental illness.

Example of a Natural Experiment

Suppose a researcher wants to test the hypothesis that poverty can cause mental illness. How can this be tested without manipulating people's income to make some people poor and some people rich? Since this would be unethical and no one would agree to participate in such an experiment, one possible alternative is to examine the consequences of some naturally occurred event that changed people's financial situation *naturally*. This is exactly what Costello and colleagues (2003), have done in their study. These researchers examined the change in the psychiatric diagnoses of 1420 children between the ages of nine and thirteen before and after the opening of a casino on an Indian reservation. After the opening of the casino, many people saw an increase in their income, and 14% of the families moved out of poverty as a result (Costello, Compton, Keeler, & Angold, 2003). Thus, this was a naturally manipulated independent variable. If the hypothesis is true that poverty can cause mental illness, the researchers ought to find a reduction in the psychiatric diagnoses of the children whose families rose out of poverty after the opening of the casino.

The main features of this study are as follows: The ***independent variable***, poverty, was *naturally* manipulated by the opening of a casino, which allowed many poor families to rise out of poverty. Thus, the experimental group (i.e., group of interest) is the formerly-poor group. The two comparison groups were the persistently poor (those whose income stayed the same) and never poor (had not been poor before and were not poor after the opening of the casino) (see Figure 6). The dependent variable was the children's psychiatric symptoms (e.g., depression, anxiety, and behavioral problems). The results of the analysis showed that behavioral problems, with the exceptions of depression and anxiety, were reduced in the formerly-poor condition after the opening of the casino. This result *partially* supported the hypothesis. But, it also raised a question of why depression and anxiety symptoms remained unaffected.

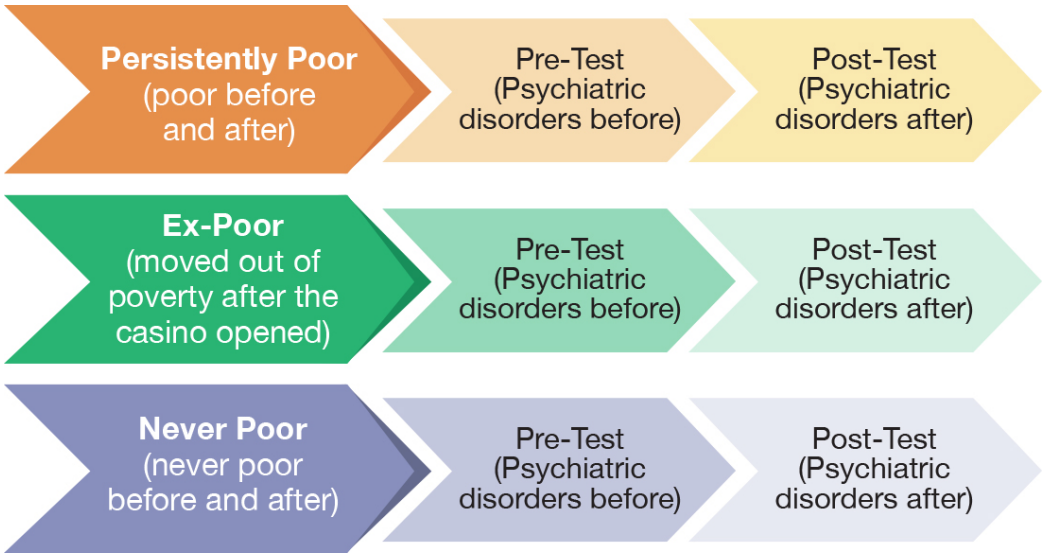


Figure 6. *A Diagram Representing the Three Groups of the Casino Natural Experiment.*

Example of a Program Evaluation

As I mentioned earlier, many program evaluations can be viewed as natural experiments, where an implemented program is the natural event *and* the independent variable, and its outcome is the dependent variable. Let's examine one example of such a program evaluation, the evaluation of a college sexual assault prevention program (Rothman & Silverman, 2007). A liberal arts college in the Northeast launched a college sexual assault prevention program in 2003, which entailed sexual assault training and education for all first-year students (see Rothman & Silverman, 2007, for more details about the program). To find out if the program was effective at reducing the number of sexual assaults on campus, the class of 2007 (these were the first undergraduates who went through the sexual prevention educational training in 2003) served as an experimental (i.e., intervention) condition. To improve the internal validity of the study, a comparison group was included, the class of 2006. These were the undergraduates who entered in 2002, one year before the program was launched. Thus, these students did not go through the sexual prevention educational training. The dependent variable was the number of sexual assaults before and during the students' college years. This was measured by an online survey, which was emailed to the students in both conditions during their sophomore years. In addition to being a natural experiment, it is also a non-equivalent control group posttest design (see Figure 7).

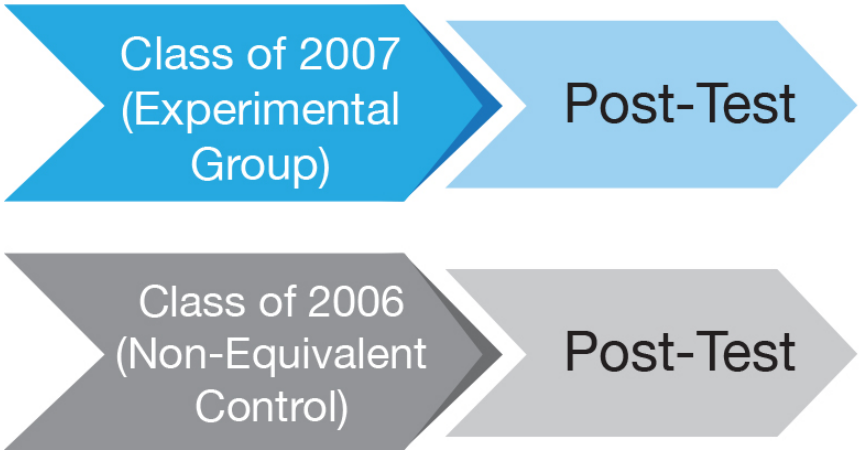


Figure 7. *A Design of Rothman & Silverman (2007) as a Non-Equivalent Control Group Posttest.*

The results of the comparison between the two groups in the reported instances of sexual assaults showed that the program was effective at sexual assault prevention, i.e., students who did not participate in the program reported more sexual assaults than their counterparts. However, the researchers also found that the program was not effective at reducing the number of sexual assaults among students with a prior history of sexual violence.

7.3 Developmental Designs

Age is one of those quasi-independent variables that cannot be manipulated experimentally. People cannot be randomly assigned to adult or child-like conditions. So, by default, developmental methods are quasi-experimental. At the same time, aging is something that happens naturally. Everyone progresses through the developmental stages of infancy, toddlerhood, early/middle childhood, and adulthood—this progression is a readily available variable. Developmental researchers recruit people of certain ages, and either follow them over time or compare them to people of different ages. The following developmental research designs employ one or both of the two techniques: cross-sectional, longitudinal, and sequential. However, just as with other forms of research methods, this is not an exhaustive list of all possible developmental methods. Let's examine each of these methods to see each of their strengths and weaknesses.

Cross-Sectional

In a cross-sectional study, people of different age groups are compared on some dependent variable. If significant group differences are found, it is implied that they are different due to their differences in age. The key feature of this design is that the comparison is made of at least two age groups. In a sense, this is a between-subjects design, where each participant is placed only in one of the age group conditions (see Figure 8).

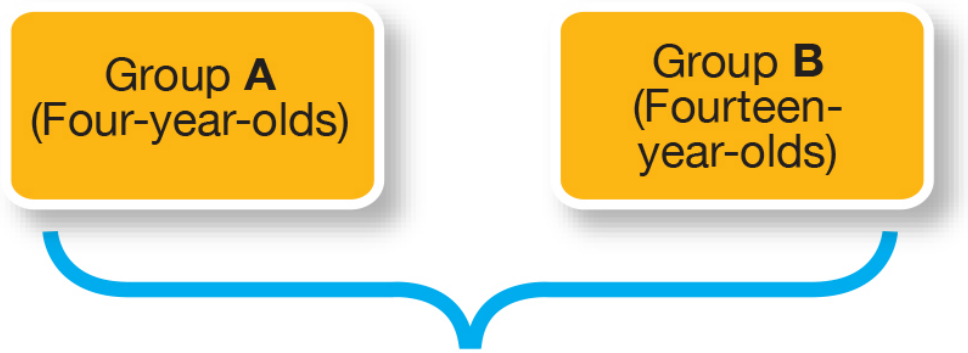


Figure 8. *A Diagram that Represents a Cross-Sectional Study comparing Two Age Groups.*

Advantages and Disadvantages

The cross-sectional design is popular among developmental psychologists because it is less time consuming, which also means that it is less expensive. In other words, a researcher wouldn't have to wait a decade for his sample of 4-year-olds to turn 14 to complete the study, he would simply use a different sample of 14-year-old children to make the comparison. However, the problems

with a cross-sectional method are almost the same as the one found in a between-subjects design. Since *different* people are tested in different conditions, any individual differences that are present between the groups, other than age, can potentially confound the outcome. But, unlike in a between-subject experiment, age cannot be manipulated in a cross-sectional study, and individual differences cannot be spread equally across the groups with randomization. This design is good to provide a snapshot of potential differences between people of different ages. However, it is virtually impossible to determine if the differences are, in fact, caused by the age differences rather than by certain other individual differences between the age groups. They are most suitable for studies that must obtain data within a short period of time and when studying certain biological differences presumably caused by aging. For example, studies on age-related differences in visual perception may employ a cross-sectional method.

A particular concern to developmental psychologists who conduct cross-sectional research is a ***cohort effect***. The cohort-effect refers to influences of historical events and/or cultural experiences that are unique to a group of people who grow up in the same time period or share the same unique experiences. These are usually generational differences. For example, imagine how different your parents' experiences were when they were your age from your own—no Internet, Facebook or Twitter, to name just a few. The unique differences caused by different cultural or generational experiences can be falsely mistaken with pure age differences. Unfortunately, a cross-sectional design cannot detect a cohort effect if in fact it has an effect on the outcome. And the simple reason is because this method does not have data on multiple generations of the same and different ages, which would require in order to separate an effect of age (pure biological maturation) from a generational effect (effect of cultural and societal differences).

The best known cohort effect, that has received a lot of attention in the field of developmental psychology, is the difference in intelligence between younger and older adults. Because most studies that have found these group differences are cross-sectional, it remains unclear whether these differences are largely due to different generational experiences, for example in educational attainment, labor market, and social programs, or due to aging (see Schaie, Willis, & Pennak, 2005).

In short, if the dependent variable under investigation is driven mainly by biological maturation such as, language or memory, then any non-age related variations that may be present between groups, such as, personality, or demographic characteristics, will probably not overly interfere with the results of the study and therefore, will not affect its internal validity. However, if hypothesized differences in the dependent variable are influenced by multiple factors, other than just by the participants' age, a longitudinal design is more appropriate to avoid the issues with confounding.

Longitudinal

In a longitudinal study, researchers follow the *same* people over time periods that can be months, decades, or even a life-span. They observe the relationships between the changes in the participants' ages and the outcome of interests such as, personality, attitudes, intelligence, memory, behavior, etc. The natural experiment by Costello and colleagues (2003), is also a longitudinal study because it followed the *same* families for 8 years (between 1993 and 2000) and made annual psychiatric assessments of their children's psychiatric symptoms (see Figure 9). Let's use this study again to discuss some of its features and issues associated with a longitudinal study.

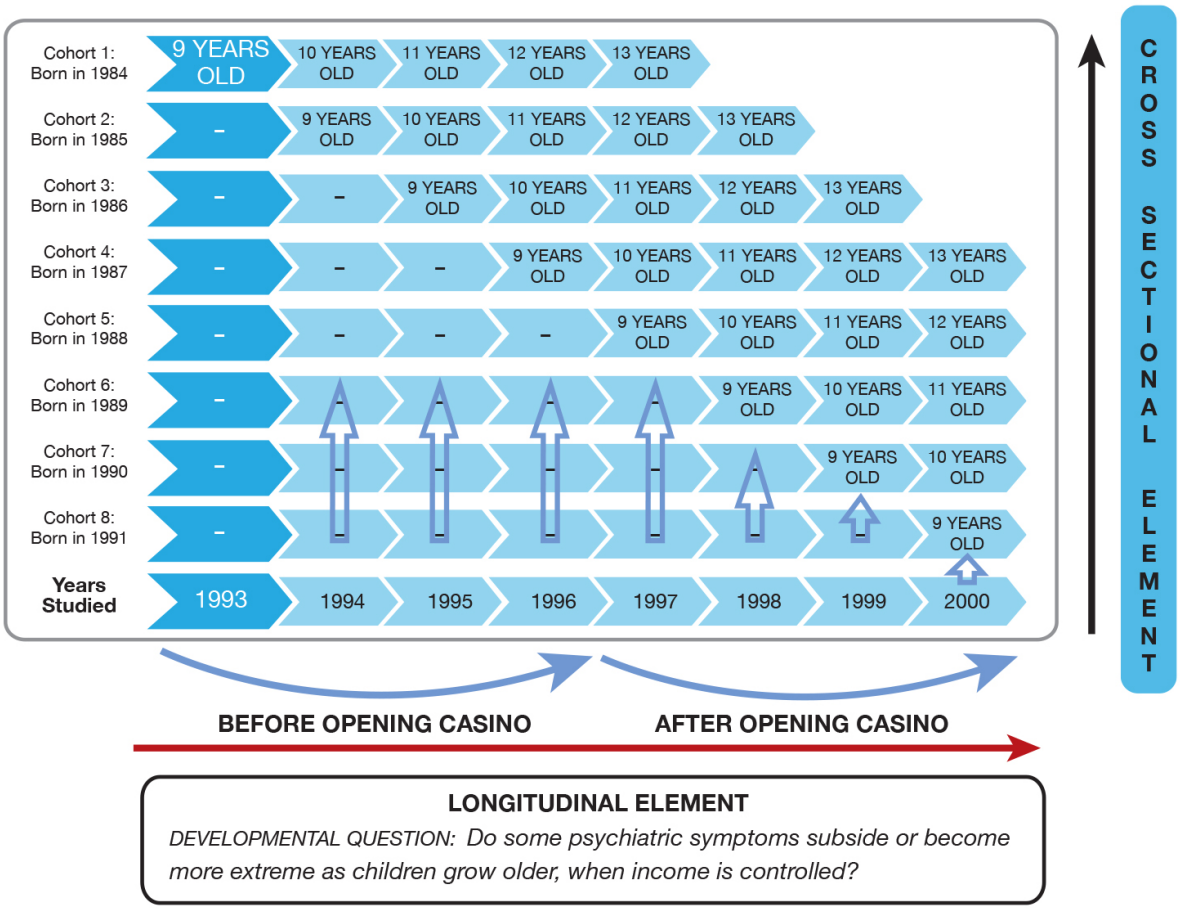


Figure 9. *A Diagram that Represents Elements of a Longitudinal and a Sequential Design using Costello et al. (2003)'s Study as an Example.*

Independent and Dependent Variables

Although age is usually one of the main independent variables in a longitudinal study, it does not have to be the only one that researchers are interested in. For example, in Costello et al. (2003),

both poverty and age were independent variables. Similarly, both could have influenced the dependent variable, psychiatric symptoms. Thus, in addition to the question about the link between poverty and mental illness, one could look at the change in the children's psychiatric symptoms as they grow older (from 9 to 13) and ask the following developmental question: Do some psychiatric symptoms subside, or even disappear, while others become more extreme as children move through childhood and into adolescence? To answer this question, one would examine the change in the number or the severity of diagnoses that each child received over the years, between 1993 and 2000, while controlling for the family income (see Figure 9; the red horizontal arrow represents the time when the measures of the dependent variable would be taken).

Advantages and Disadvantages of a Longitudinal Study

As the study by Costello et al. (2003), demonstrates, the longitudinal method yields rich data and corrects the issues associated with participants' individual differences. But, it does have its disadvantages. First, it can take years or decades to complete one study, so it can be very expensive. Some researchers do not have the time or such resources. Other problems can arise from repeatedly testing the same people, as in a, within-subjects design. While fatigue may not be as relevant in a longitudinal study, since the time between assessments are more spaced, practice and history effects can still be a threat to the internal validity of a longitudinal study.

Cohort effect can also be a problem for a longitudinal study, mostly by reducing its external validity. If a researcher studies one particular cohort of people, it may be questionable whether the results can apply to people of other generations or experiences. Twenge and Campbell (2008), reported the following interesting generational differences between young people of this and past generations. College students of today are more narcissistic and have higher self-esteem. They have low social approval need and are more likely to attribute causes of events to external factors other than to themselves. More Americans are treated for major depression and anxiety, even when compared to adults who lived through the Great Depression. And, more women are now in the workforce and score higher on assertiveness and all other listed traits. It is no wonder that the current generation has been labeled "Generation Me". We simply don't know yet whether research on the, "Generation Me" cohort, speaks only to this unique generation or whether it can tell us something about all people in general.

Cohort effects are not always generational, some operate on just a particular group of people because of their unique experiences. For example, Carmil and Breznitz (1991), found that Holocaust survivors *and* their children shared similar political attitudes, religious identity, and future orientation and that they were different from the control group of non-Holocaust survivors or their descendants. Specifically, Holocaust survivors and their children supported more centrally oriented political parties, expressed a stronger belief in God and had a more positive outlook about the future.

Of course, cultural or historical experiences do not have to be extreme to have an impact on people's attitudes, beliefs, opinions, and behavior. We are all influenced by our era in some way or another. So, to control and even study cohort effects, researchers can include multiple cohorts into one longitudinal study. They can then compare them on the same dependent variable. The logic of this method is similar to adding a control group to control for history effect in a within-subjects design. This allows researchers to spot and measure any cohort-related differences and separate them from the developmental changes. A longitudinal study that combines multiple cohorts into one longitudinal study is known as a ***sequential design***.

Sequential Design

As was already mentioned, the sequential design is nothing more than a longitudinal design but, with multiple cohorts followed over time. When two or more cohorts are added to a longitudinal design, the study takes on the elements of a cross-sectional and a longitudinal design. Figure 9 illustrates this point using the Casino study (Costello et al., 2003), as an example. The horizontal red arrow is the longitudinal part of the study (the same families and their children are followed over a period of 8 years), and the rows represent the cross-sectional part of the study, these are different cohorts and different age groups. The black arrows further demonstrate that when two or more cohorts, of different age groups, are compared, it resembles a cross-sectional design and will render the same information as one would obtain from a cross-sectional design study. For example, if the study is in its 5th year (i.e., 1997), the 9-year-olds of cohort 5 can be compared with the 10-year-olds from cohort 4, the 11-year-olds from cohort 3, and so on (see Figure 9). But in addition, this design will also provide the longitudinal information, the changes that occur within each individual and each cohort over time. Finally, this design can help researchers uncover any hidden cohort effects and draw more accurate conclusions about developmental influences.

Returning to the questions about age-related differences in intelligence, systematic comparisons between different cohorts of the same ages and examining the trends in intellectual changes within cohorts can address them. To give an example, Schaie, Willis, and Pennak (2005), have uncovered some interesting patterns in intelligence by examining the Seattle Longitudinal Study (SLS) data, which contain 13 birth cohorts, born between 1889 to 1973, followed between 1963 and 1998 in 7-year intervals. Overall, they confirmed the generational trend of upward increase in *fluid* (e.g., inductive and spatial reasoning) and *crystallized* (verbal and number ability) intelligence since 1900s (i.e., cohort effect). These positive gains were attributed to multiple social and economic factors such as, public policies related to child labor laws, educational regulations starting with 1900s, federal funding of education, changes in educational curriculum and pedagogy, and an increase of women in the workforce. These changes are also directly linked to the changes in the societal attitude towards the value of education (Schaie et al., 2005). But, they also found some interesting negative trends within some cohorts. For example, the following trends were found

within the Baby Boomer cohort (1945-1966): an *increase* in inductive reasoning, a *decrease* in number ability, and, no change in verbal ability.

References

Carmil, D., & Breznitz, S. (1991). Personal trauma and world view—are extremely stressful experiences related to political attitudes, religious beliefs, and future orientation. *Journal of Traumatic Stress*, 4, 393-405.

Costello, E. J., Compton, S. N., Keelers, G., & Angold, A. (2003). Relationships between poverty and psychopathology: A natural experiment. *The Journal of the American Medical Association*, 15, 2023-2029.

Schaie, K. W., Willis, S., L., & Pennak, S. (2005). A historical framework for cohort differences in intelligence. *Research in Human Development*, 2, 43-67.

Twenge, J. M., & Campbell, S. M. (2008). Generational differences in psychological traits and their impact on the workplace. *Journal of Managerial Psychology*, 23, 862-877.

Chapter 8: Understanding the Logic of Statistics and Describing Data Like a Scientist

Introduction

Up to this point we have been discussing different methods of ***data collection*** that address different research questions. But in this and the following chapter, we are going to look at the next phase of quantitative research, the ***statistical data analysis***. This is the stage in the research process when a hypothesis can be finally tested. This process is based on the logic and mathematical computations of statistics.

Why do social scientists use statistics to test their hypotheses, you might ask? After all, statistics is a separate discipline, that concerns the development and studies methods of collecting, analyzing and interpreting empirical data. But suppose I want to test a theory, that smiling in humans has evolved to promote social connections and alliances. To test it, I predict that more smiling should occur in larger than in smaller groups since more effort is needed to build cooperation in groups of more people.

Suppose I observe one small and one large groups of people, and count the number of smiles people displayed in each group during a one- hour interaction. Assuming my observation revealed that people in the large group smiled 10 times more than people in the small group, can I conclude that my theory was correct? Was the difference big enough to support my hypothesis? The answer is we don't know. And here is why: the uncertainty comes from the fact that I did not observe all different kinds of groups and people—I relied on a small sample of some people of certain characteristics and from only one place. Given this significant drawback, I will use the strategies of Statistics to circumvent the issue of sampling in order to make sense of my results.

In this chapter, we will touch on the logic and principles of the probability theory that is at the foundation of statistical testing. But we will first begin with descriptive statistics.

Describing and Summarizing the Sample

Before a researcher can, statistically, test any hypothesis the data should be described and summarized to make sure that its properties meet the researcher's expectations. Descriptive statistics are also useful and should be used to look for any hidden or unusual patterns in the scores, as well as to check for errors or typos when entering data on a computer. The two most commonly used types of descriptive statics are measures of ***central tendency*** and ***dispersion***. But even before

measuring the central tendency and the dispersion of your data, it is always a good idea to do a visual examination of its distribution.

Data Distribution (Histograms and Bar Charts)

The quickest and the easiest way to get a feel for your data is to graph and examine the distribution of the values in order to see how frequently each score appears in the data. This can also reveal any unusual or extreme values in the data, which could be either an error or an actual obtained value. One of the common graphs used is a ***histogram***. This graph is used to represent a ***frequency distribution*** of an interval or a ratio scale variable (see Graph 8.1).

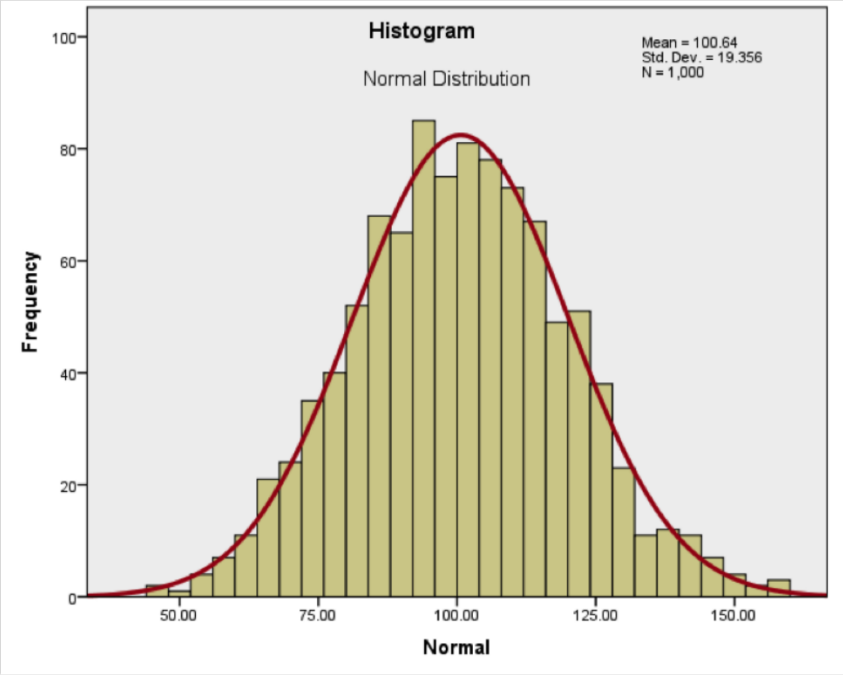


Figure 8.1. *Example of a Histogram.*

The x-axis represents values in a set of data for a particular variable and is divided into bins of equal width. The number of data points (individual scores) that sit inside each bin are represented by the bars whose heights are the frequencies, or the number of the scores, that occur in each bin. Thus, the data set that is displayed in Figure 6.1 has 1,000 scores (individual data points) but each score appears only once; there is only one score of 46.51, one score of 46.58, etc. So, in this case it is especially helpful and more informative to use a histogram to visualize distribution by grouping the scores into bins and looking at the frequencies of the groups of scores.

Another important reason to graph your data before conducting any statistical analysis is to know whether it is ***normally distributed***. This is an important assumption on which most ***parametric*** inferential statistics are based, a subject covered later in this and subsequent chapters. Normally distributed data has a bell-shaped curve, just like the bell curve of the histogram on Figure 8.1. Notice that normally distributed data is also ***symmetrical***, which is one easy way to visually determine whether data is normally, or not normally, distributed. We will come back to the properties of normally distributed data when we cover statistical averages and standard deviation next.

To graphically represent the frequency distribution of a nominal or an ordinal variable, use a ***bar chart***. The bars represent the categories or the ranks of a given variable. Their heights correspond to the frequencies of their occurrence in the data. For example, take a look at the bar chart in Figure 8.2. The represented nominal variable has two gender categories: male and female. The heights of the two bars represent the frequency of the male and the female genders in this hypothetical sample. The graph illustrates that there are 12 female and 8 male participants in this sample.

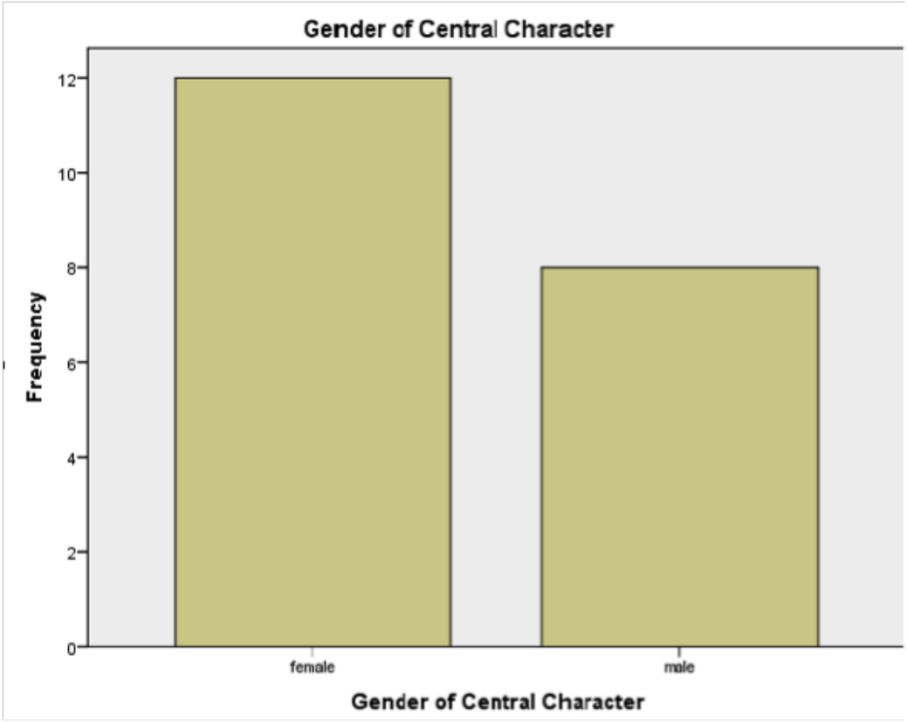


Figure 8.2. *Bart Chart of a Nominal Variable.*

To graph a bar chart or a histogram using SPSS, follow these steps: Go to Analyze →Descriptive Statistics→Frequencies→select the variable you want to graph (e.g., Gender of Central Character)→move it to Variable(s) box→select Charts option and select bar charts or histogram→Continue and OK (see Figures 8.3 and 8.4).

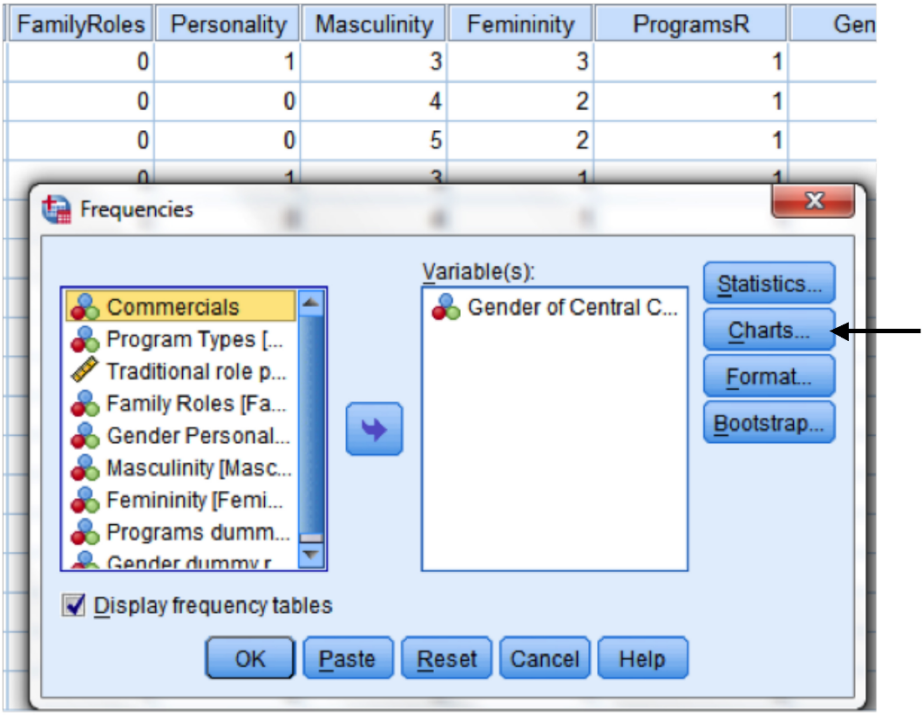


Figure 8.3. *Making a Graph.*

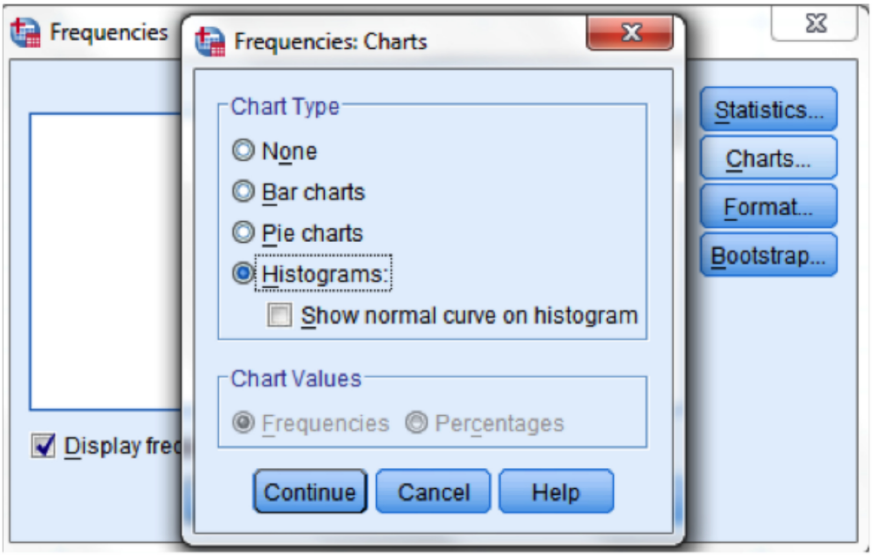


Figure 8.4.

Measures of Central Tendency (i.e., Averages)

Central tendency is the most representative value in a given set of scores and it happens to be located around the center of a normally distributed dataset, i.e., *central tendency*. If you re-examine

the histogram in Figure 8.1, you will notice that the most frequently occurring values are grouped around the score of 100, which happens also to be their centrally-located average. The usefulness of central tendency is that this single value can be used to describe an entire dataset. For example, knowing that the average of 1,000 scores is 100 gives the sense of how big or small most of those values are.

There are three types of measures of central tendency, also known as averages: ***mean***, ***median***, and ***mode***. In a perfectly normal distribution all three averages are the same. However, as the distribution gets more skewed (becomes less symmetrical) the mean becomes the most affected as it gets pulled by the extreme values in the distribution. Thus, the mean should be used for normally distributed data only. If data is skewed, the median is a better alternative. Another consideration to make when choosing an average to describe central tendency is the scale of the variable. The mean should be used for a ratio and an interval variable. Median is a better choice for an ordinal or a skewed interval/ratio variable, while mode is useful to describe nominal data. Finally, the choice to use one particular measure of central tendency over others may depend on the purpose that it will serve.

Mean

Mean is a mathematically computed average of a dataset. It is derived by adding all values in the set and dividing them by the number of scores in that set. For example, the mean of the scores: 5, 10, 3, 7, 6 is computed by (5+10+3+7+6)/5, which is equal to **6.2**. Because a mean describes the distance of a given score from that mean, it should be used to describe the central tendency of a variable measured on a *ratio* or an *interval* scale only, since the distance between the points (or ranks) in these scales are equally spaced and, thus, yield meaningful information. However, when a distribution contains extreme values, a.k.a.*outliers*, computing its mean may not be the best representation of that distribution. For example, suppose I administered a test to 10 students and my test is a ratio scale variable. The scores can range between 0 and 100. Suppose, that I obtained the following distribution of the exam scores (see Table 8.1). Note, that student #4 and #10 never took the test, accounting for the, 0s.

Table 8.1. Example of a Distribution of Exam Scores

Students	Exam Points
1	81
2	85
3	90
4	0
5	83
6	91
7	80
8	79
9	95
10	0

The mean of the exam scores is 68.4 The mean score doesn't seem to be a fairly representative value because, all but two students earned a much higher grade. The reason why the mean turned out to be lower than it should have been is because of the two students who never took the test and have zero points, these are the outliers. They are unusually low scores compared to the rest of the scores in the distribution. There are a couple of options to deal with such outliers. One, the outliers can be omitted from the calculation of the mean altogether, but only if a researcher can justify their omission. This is why you need to know the purpose of calculating the measure of central tendency. The purpose of computing the mean of the exam scores is to find out how well or how poorly the students did on the exam. Since the zeros reflect the missing scores, rather than the quality of the two students' work, their omission is justified. The second option is to calculate the ***median*** *instead*

*of the **mean***. This would be a better option if the zeros were the failed scores and their inclusion in the calculation would be necessary to account for the failures. Generally speaking, extreme low scores will pull the mean down, decreasing it significantly and extreme high scores will pull the mean up, by significantly increasing it. Thus, if you want to get an accurate estimate of all the scores and your data is skewed by one or several extreme scores, you should compute a median.

Median

Median is a value that is the midpoint in a distribution of scores, so that there is an equal number of scores above and below it. To calculate the median of the exam scores used earlier (see Table 8.1), first I need to order the scores from lowest to highest and then find the score that is in the middle of that distribution. Thus, my exam distribution should look like this: 0, 0, 79, 80, 81, 83, 85, 90, 91, 95. However, because it has an even number of scores, there aren't one but two middle scores: 81 and 83. To derive one midpoint I will have to compute the mean of the two middle scores: (81+83)/2—which equals **82**. Right away, it is clear that this is a more accurate representation of the exam scores than the mean of **68.4**. This is exactly why the median is oftentimes used as a representative of a data set that contains extreme values.

Another instance when a median and not a mean should be used to calculate central tendency is when data are measured on an *ordinal scale*. If you recall, ordinal data contain information about the direction or the ranking, but the spacing between the points or the ranks is not equivalent. Another way to put it is to say that a scale of 1, 2, and 3, for instance, can be also represented as 1, 5, and 7 because the values only reflect the order that the value of '1' represents the first place, the value of '5', second, and '7 represents the third/last place in the ranking. But, the values themselves do not reflect the actual magnitude or difference between the ranks. Thus, many consider it inappropriate to compute a mean of an ordinal distribution, and instead, use a median as the measure of central tendency. Still others believe that it is acceptable to compute a mean of ordinal data as long as it is appropriately interpreted.

Mode

Mode is the most frequently occurring score in a given distribution of scores. Thus, the mode of my earlier exam grades' distribution is zero, because, unlike the other values that only occurred once, zero occurred twice. This was, clearly, not the most accurate representation of the exam grades. But mode can be useful in other cases. It may also be used to calculate the central tendency of a nominal variable and to represent the most frequent category in a dataset.

To illustrate this point, suppose a researcher is interested to know the type of bullying that most frequently occurs in public schools. To do so, he categorizes observed instances of school bullying into two types, psychological and physical, and then computes *the mode*, the most frequently observed category.

Measures of Dispersion

This descriptive statistic tells us how spread out or *dispersed* the scores are in a given distribution, which also speaks to the amount of variability in the dataset. Perhaps the easiest measure of dispersion to understand and to calculate is a ***range***, the difference between the largest and smallest values in the distribution. The problem with using range as a measure of dispersion is that it is sensitive to extreme scores (i.e., outliers) just like a mean, and so may not be an accurate representation of the dispersion in a distribution that contains extreme values. The most commonly used measures of dispersion in both, descriptive and inferential statistics, are ***variance*** and ***standard deviation***.

Variance and Standard Deviation

Both measures of dispersion tell us about the average distance of the scores in the distribution from the mean. The symbol for population variance is σ^2 (sigma squared), and 2^2 for a sample variance. To compute sample variance, *first find the distance between the mean and every score in the distribution, square each value, deduce the sum, and divide the sum by the number of scores minus one. We square the deviations in order to avoid a zero sum when we add all the deviations.* For example, suppose we have a distribution of the following three scores: 1, 2, and 3, which has the mean of 2. If we add all three scores' deviations without squaring them, we get the following answer: (1-2) + (2-2) + (3-2) = (-1) + (0) + (1) → 0. So, to avoid the zero sum, we have to square each deviation first. Doing so, we get the following answer→ $(-1)^2 + (0)^2 + (1)^2 \rightarrow 1 + 0 + 1 = 2$. The final step is to divide the sum of the deviations by the number of scores minus 1→ 2/(3-1), and we get the following variance: $s^2 = 1$.

The ***standard deviation*** is simply the square root of the variance. For example, the standard deviation of 1 (from the previous example) is $\sqrt{1}=1$. We calculate that in order to correct for the squaring of the deviations. The symbols for the population and the sample standard deviation are σ (sigma) and *s*, respectively.

The good news is that most, if not all, statistical packages (including SPSS) will easily compute both, a variance and a standard deviation, for you so, almost no researcher has to calculate them by hand.

To further illustrate how informative this descriptive statistic can be, take a look at Figures 8.5 and 8.6, which display two hypothetical distributions of scores. Note that the means of both distributions are very similar, and both have an equal amount of cases. Yet, they don't look very much alike—the dispersion of the first data set is smaller than that of the second data set and their standard deviations reflect that difference, $s_1 = 10.1$ and $s_2 = 16.7$, respectively.

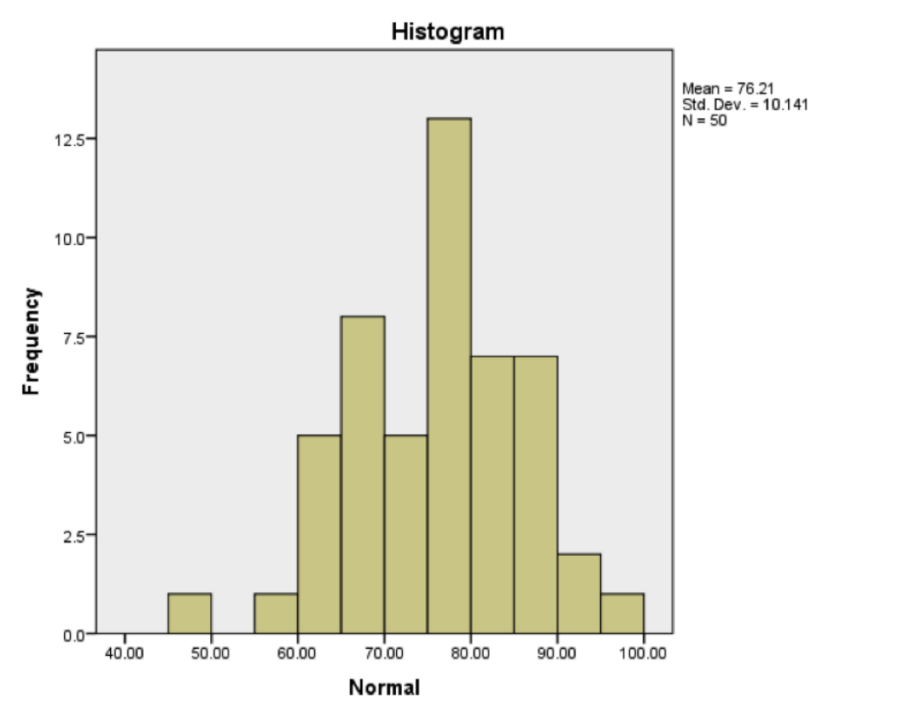


Figure 8.5. *Distribution of Data Set 1.*

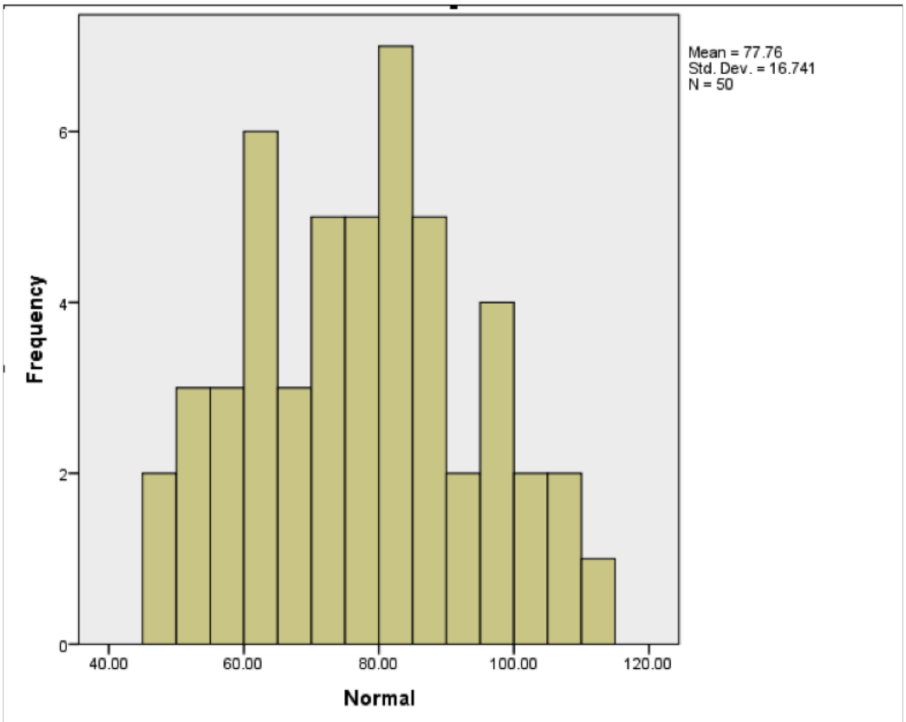


Figure 8.6. *Distribution of Data Set 2.*

Skewed/Not Normally Distributed Data

Not all data can be normally distributed or symmetrical. Sometimes data can be skewed to the left, when the tail is on the left, or to the right, when its tail is on the right. A histogram can usually reveal such patterns. As examples, I graphed three histograms of three different quizzes from my Research Methods class. Let's take a look at each one of them.

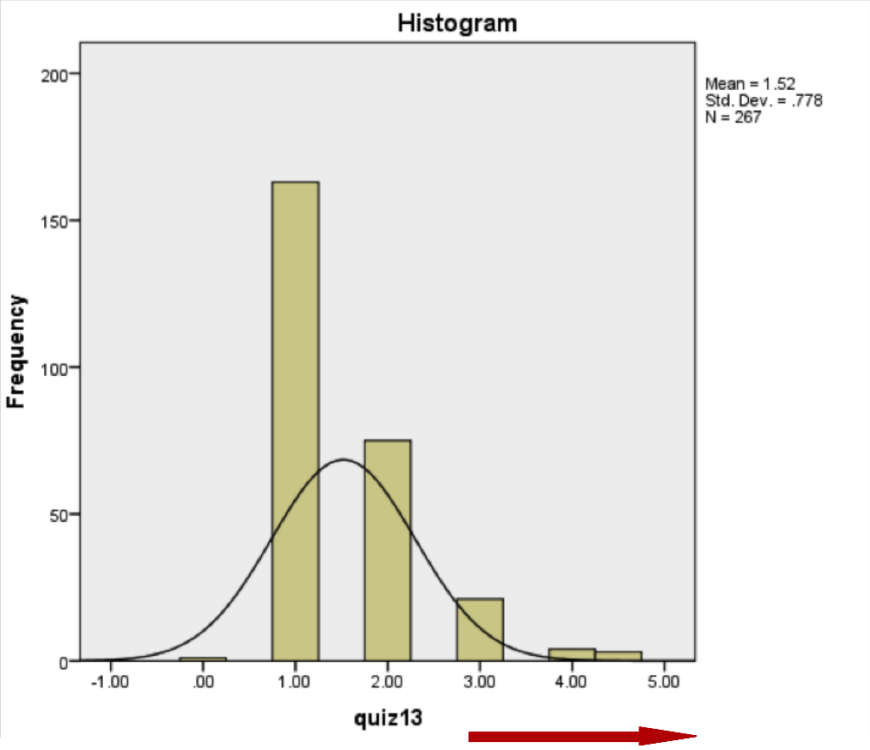


Figure 8.7. *Quiz 13 Grades Skewed to the Right.*

Note that each quiz has a total of twelve questions, which means that each student can get a minimum of zero or a maximum of twelve correct answers. The histogram depicted in Figure 6.7 displays the frequency distribution of quiz 13, skewed to the right. What this distribution shows is that the majority of the students only got 1 (about 150 students) or 2 (another 100 students) answers correct. The histogram also reflects that very few got five or four questions correct (less than 50), and no one answered more than five questions correctly. This must have been one tough quiz!

Remember that the mean will be affected if the data is skewed. In this case, when it is skewed to the right, the mean tends to be greater than the median.

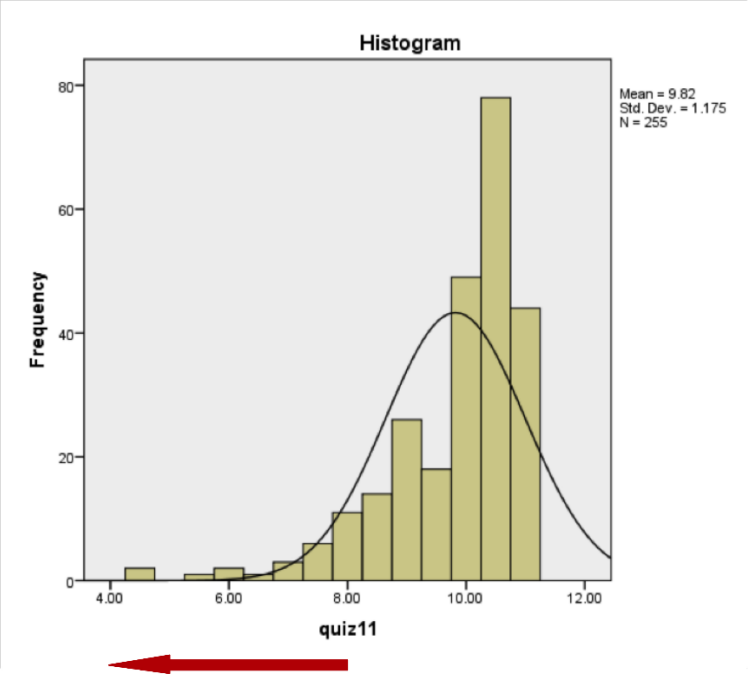


Figure 8.8. *Grade 11 Grades Skewed to the Left.*

On the other hand, quiz 11 seems to be an easy one! The histogram represented in Figure 8.8 shows that it is skewed to the left of the distribution, with most of the scores grouped around ten and eleven. Most students were able to answer at least ten or eleven questions on this quiz. Also, the mean of the skewed to the left data tends to be smaller than the median.

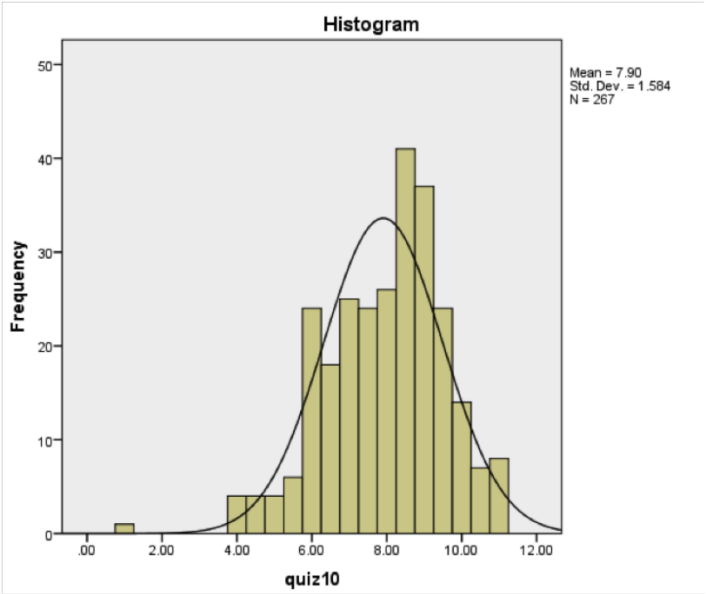


Figure 8.9.*Quiz 10 Grades Normally Distributed.*

Quiz 10, on the other hand, seems to have a good balance of grades that are both below and above the mean of **7.90**. This shows that the average student answered about seven to eight questions out of the twelve available, and that a relatively equal number of students answered below and above that number. Although it is not a perfectly symmetrical distribution it does appear to have a bell-curve shape. Thus, this distribution is normal. You may recall that the mean, the median, and the mode of a normal distribution are approximately the same, which is what we can see in this case if we compute all three averages—the median is **8**, i.e., the midpoint in the distribution, and the mode is 8.5, the most commonly occurred number in a distribution, and the mean is 7.90.

One final characteristic of a normal distribution: about 99.7% of the data tends to lie within three standard deviations of the mean, about 95% lies within two standard deviations, and at least 68% of the data should be located within one standard deviation of the mean. This is known as the ***empirical rule*** and pertains to all normally distributed data. So we can apply this empirical rule to quiz 10 data and predict that about 68% of the students who took that quiz got between 6 (7.9 - 1.58) and 10 (7.9 + 1.58) answers right, and about 95% of the students answered at least 4 (7.9 – (2 * 1.58)) and 12 (7.9 + (2 * 1.58)), given that the *standard deviation* of the distribution is 1.58.

Correlations

If data contain two or more variables, knowing whether they are associated with each other is very informative. It is measured statistically, by computing a ***correlation***, a statistical measure of association. It is typically easily accomplished with any statistical software, e.g., SPSS. Correlation, however, does not mean ***causation***—that one of the correlated variables is the cause of the other. Cause and effect relationship can only be tested experimentally using an experimental method.

Two pieces of information a correlational analysis will give you and you want to know: the strength and the direction of the relationship. In terms of the strength, the value of the correlation coefficient can range between -1 to +1. A value of [1] indicates a perfect, 100% association between two variables; whereas 0 indicates a complete absence of the association. The closer the value to 0, the weaker the association. A sign -+ denotes the direction of the correlation, positive or negative.

Positive Correlation

A correlation is said to be *positive* when the values for both variables increase or decrease *together*, i.e., both variables change in the same direction. For example, physical exercise and health are positively correlated. When we exercise more, i.e., ***increased*** exercise, our health tends to improve as well (***increased*** good health). More studying, i.e., ***increased*** time for studying, before taking a test is supposed to be positively correlated with the test score, i.e., ***increased*** test score.

Alternatively, we can say that as physical exercise is ***decreased*** so does health. When students ***decrease*** the amount of time studying for a test their test grades are likely to ***decrease*** as well. Whether the values decrease or increase, they change in the same direction. This type of association can be seen graphically. Figure 8.10 represents a positive correlation between variable 1 (x-axis) and variable 2 (y-axis). The ascending line with a ***positive*** slope is the best fitted line and reflects the direction of change in the values of the two variables.

Negative Correlation

When two variables change in the opposite directions, the value of one variable is increasing while the value of another variable is decreasing, the correlation between the two is said to be ***negative***. For example, there is a negative correlation between poverty and health. People who live in poverty (experience more economic hardship) have poorer (lower) physical and mental health. Negative correlation should not be interpreted as a "bad" association, only that the values of the variables go in opposite directions or that the slope of the line representing the association between the variables is negative (see Figure 8.11). We know that exercising is good for one's health because it can reduce excess weight. Thus, the correlation between exercise and weight is usually negative.

No Correlation

When no relationship exists between two variables and the correlation is 0, the line will look flat, just like the one in Figure 8.12

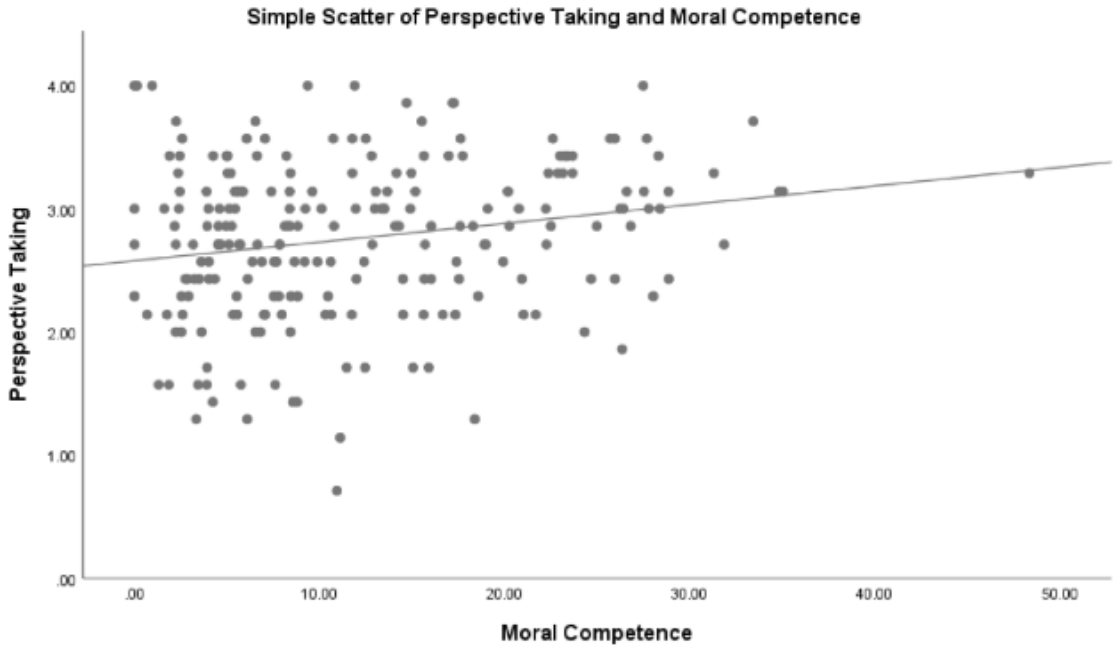


Figure 8.10.*The Positive Slope of the Line Represents the Positive Correlation Between Perspective Taking and Moral Competence.*

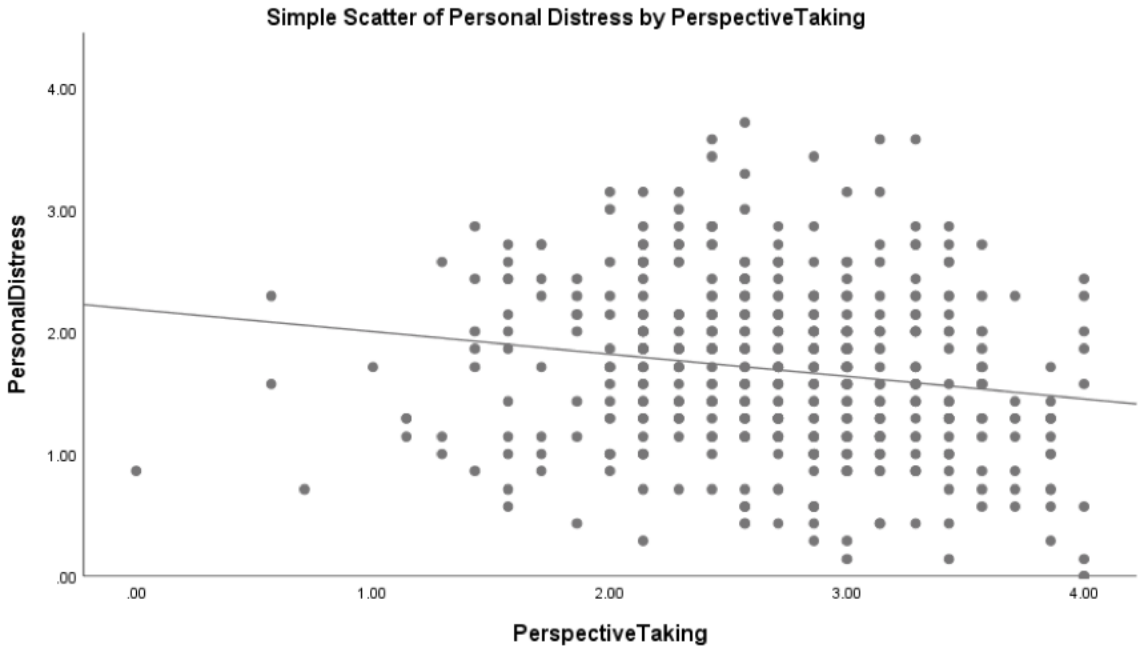


Figure 8.11.*The Negative Slope of the Line Represents the Negative Correlation Between Personal Distress and Perspective Taking.*



Figure 8.12. *No Correlation between Personal Distress and Moral Competence.*

8.2 The Logic of Inferential Statistics in Hypothesis Testing

Descriptive statistics are used to describe and to summarize data, however, they do not permit researchers to make any conclusions about the "correctness" of a hypothesis. To test a hypothesis, we use ***inferential*** statistics. The term comes from the fact that we draw ***inferences*** about a population of interest from the results of our sample. But before we can move to the topic of computing inferential statistics, let's take a look at the logic of this process so you can better appreciate and understand how to interpret results of an inferential statistical analysis.

Theory of Probability

Suppose you conducted a perfectly designed and executed experiment to test a hypothesis that children would display more aggressive acts after they played an action video game. Let's further imagine that, after you conducted this experiment, the means of aggressive acts observed before and after were seven and ten, respectively. Can you conclude that your hypothesis was correct given the difference in the aggressive behavior before and after? Not yet. There are several problems with jumping to a conclusion from simply looking at the mean differences. One of the biggest problems is the fact that you did not observe all children in the ***population*** to test your hypothesis. And since you didn't there is a possibility that the obtained results were obtained by chance only. In other words, can you trust the results from a sample to apply them to the population? We don't know unless we conduct an inferential statistical analysis.

There are, of course, many other uncertainties that may prevent the researcher from generalizing their findings to the population of interest.

"What is the probability of obtaining the results, as extreme as from the given sample, if the null hypothesis were true?"

Since in most cases, researchers are unable to test an entire population of interest, and because there are always going to be many uncertainties about a phenomenon in question, even in a perfectly controlled experiment, inferential statistics use the logic and principles of the probability theory to test hypotheses by determining the probability of obtaining the results of a study, as extreme, if the null hypothesis (the opposite of what the researcher predicted) is true. In other words, the only way that we can "prove"[5] that our hypothesis is right is by calculating the

[5] In science we can't ever prove anything definitely. So, instead of using the word *prove*, social scientists will say *support*. I used the word *prove* to highlight the intention of what we want to do when we test a hypothesis, rather than what we can actually achieve.

probability of being wrong (that there is no effect) and still obtaining the results (the results can be suggesting that the tested hypothesis is correct). Let's take a look at how this logic is used in hypothesis testing statistically.

Statistical Testing

Inferential statistical analysis begins by stating one's hypothesis as a null (H_0) and an alternative hypothesis (H_1). A null hypothesis assumes that no difference or no effect exists; it is, essentially, a statement that the researcher's hypothesis is wrong. I like the analogy of presumption of innocence in the court when judging a defendant. The problem is similar in that it is not possible to prove the guilt and so the judge must consider the evidence that will lead them to reject the presumption of innocence. If the innocence is rejected, then the guilt is assumed by default. Alternatively, if rejecting the innocence is not possible, the guilt cannot be determined. Statistical hypothesis testing uses the same logic. An alternative hypothesis is the one the researcher is interested to support. (Or it is an attempt to reject the innocence of the defendant in the analogy with the court room). But it is the null hypothesis that is tested by computing one of the inferential statistical analyses. This test will produce a statistic (a number) and the probability ***(p-value)*** of obtaining the results, as extreme, if the null hypothesis were to be true. This probability can range between zero, no chance, and one, 100% chance. In social sciences, it is acceptable to have some chance of an error. The convention is to allow a p-value of .05 alpha level (α) or less, which means that if the null hypothesis were true (and the researcher's hypothesis is false) a result of this sample this extreme would occur 5% of the time. This is accepted as small enough to reject the null, and by default, to accept the alternative hypothesis. Keep in mind that virtually all statistical packages will calculate these for you. All you need to know is how to interpret the output, which gives you the results of the statistical analysis computed by the software.

An Interesting Note: Notice the similarity in the logic of falsifiability and statistical testing! According to the principle of falsifiability, a scientific claim (a theory or a hypothesis) should be tested by trying to disconfirm it. In statistical testing, we use the same sort of principle. Researchers test the null hypothesis by assuming that no difference or true effect exists and hope that they can disconfirm it.

Type I and Type II Errors

The process of statistical hypothesis testing is not perfect and allows for errors in judgment. So, when a scientist makes a decision about whether to accept or reject a null hypothesis, two possible types of errors can be committed: a type I or a type II error. A Type I error is said to occur when we reject the null hypothesis, when in fact, there is no true effect or differences between the groups. In other words, it is when a research hypothesis is not true but it is declared to be true. Much like

a *false positive*. A researcher can have some control over committing a type I error by setting α level as small as needed, 0.05 alpha level or even less, depending on the seriousness of the consequences of committing a type I error.

A Type II error is committed when a researcher fails to reject a null hypothesis, when in fact, true effect or group difference was found. As when a research hypothesis is true, but it is declared not to be true, as a *false negative*.

Notice that if we reduce the probability of committing a type I error, e.g., by lowering the alpha level standard (e.g., lower than the standard of 0.05), we are increasing the probability of committing a type II error because setting a p-value to a lower alpha level makes it difficult to find significance and to reject the null hypothesis. To reconcile this, a researcher should decide which would be more consequential—if he/she commits a type I or type II error.

For example, suppose you have developed a new drug and you want to make sure that it is safe. First, let's establish your null and your alternative hypotheses:

H_0: No association between taking the new drug and dangerous side effects (No association exists).

H_1: There is an association between taking the new drug and having dangerous side effects (Hopefully, this is not the case.)

Type I error: Occurs when the researcher rejects the null hypothesis and assumes that there are dangerous side effects, when in fact it is not true.

Type II error: Occurs when the researcher does not reject the null hypothesis and assumes that the drug is safe (no association between the drug and dangerous side effects), when in fact, it is (the drug is not safe).

If the drug is a new cold medicine, it is probably more desirable to commit a Type 1 error (to assume that it is unsafe) rather than a Type II error. In other words, you would rather be extra precautious and remove the new drug from the market (rather than risking to poison many people when there are other safe cold medications already available on the market). On the other hand, let us imagine that this new drug is a potential treatment for the Alzheimer's disease. Considering that Alzheimer's patients haven't had any viable treatments, they might be more inclined to take their chances and try this new drug. So in this case, it might be more acceptable to increase the chances of committing a Type II error so that the new treatment drug can quickly reach the patients.

Later in the book we will discuss a growing concern among social scientists that many published results cannot be replicated, which may be due to frequent occurrence of a type I error. In other words, an argument can be made that studies of social nature, that may not have serious

consequences (although you may argue that that is also not true) if a type I error is committed, may be too casual in their reliance on the use of .05 level of significance.

Computing Correlations with SPSS

Now that you are familiar with the statistical hypothesis testing, let's compute and interpret a few correlations using real General Social Survey data **(GSS).**[6] **This is an ongoing research project which was started in 1977** by the Social Science Research Center at the University of Chicago, and supported by the National Science Foundation. I will use the 2018 survey to demonstrate the steps of computing correlations. **Important:** *Pearson product-moment correlation coefficient*is a statistical measure of association to be computed only if both variables are interval or ratio (continuous).

Begin with:

Analyze→Correlate→ select **Bivariate** (see Figure 8.13)

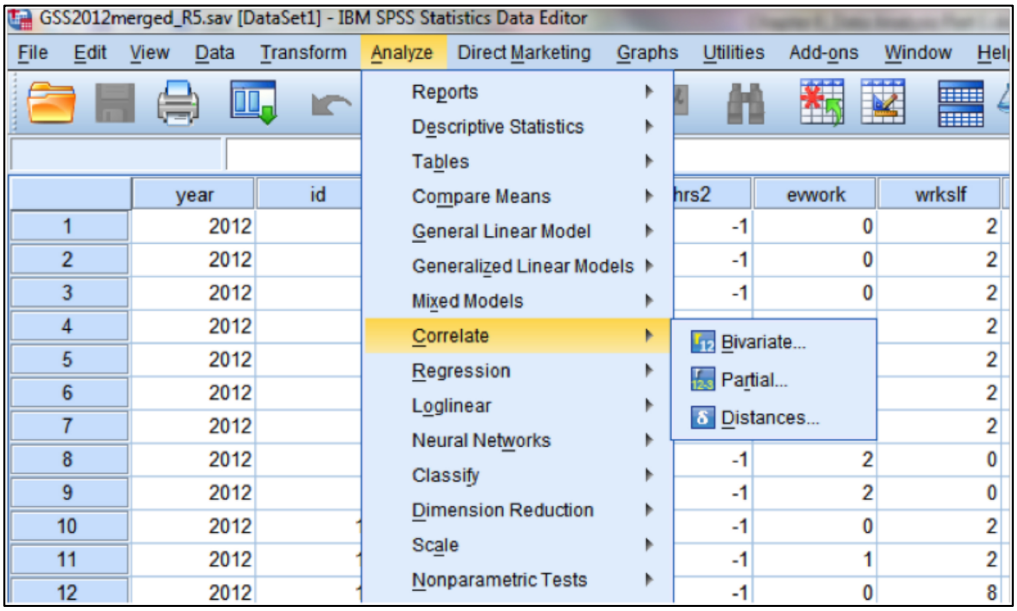


Figure 8.13. *Computing Correlations.*

[6] To visit GSS website, follow this link http://www3.norc.org/GSS+Website/Download/SPSS+Format/; select 2018 from **Individual Year Data Sets (cross-section only)**.

From the drop down menu, select the variables you want to correlate with each other and move them to the *Variables* box. In my example, I would like to find out if there is a correlation between the respondents' income and the hours per day they watch TV. Thus, I selected these two specific variables and moved them into Variables box. Select Pearson as your correlation coefficient and *Flag significant correlations* (all statistically significant correlations will be flagged at the p value of less than 0.05) (see Figure 8.14).

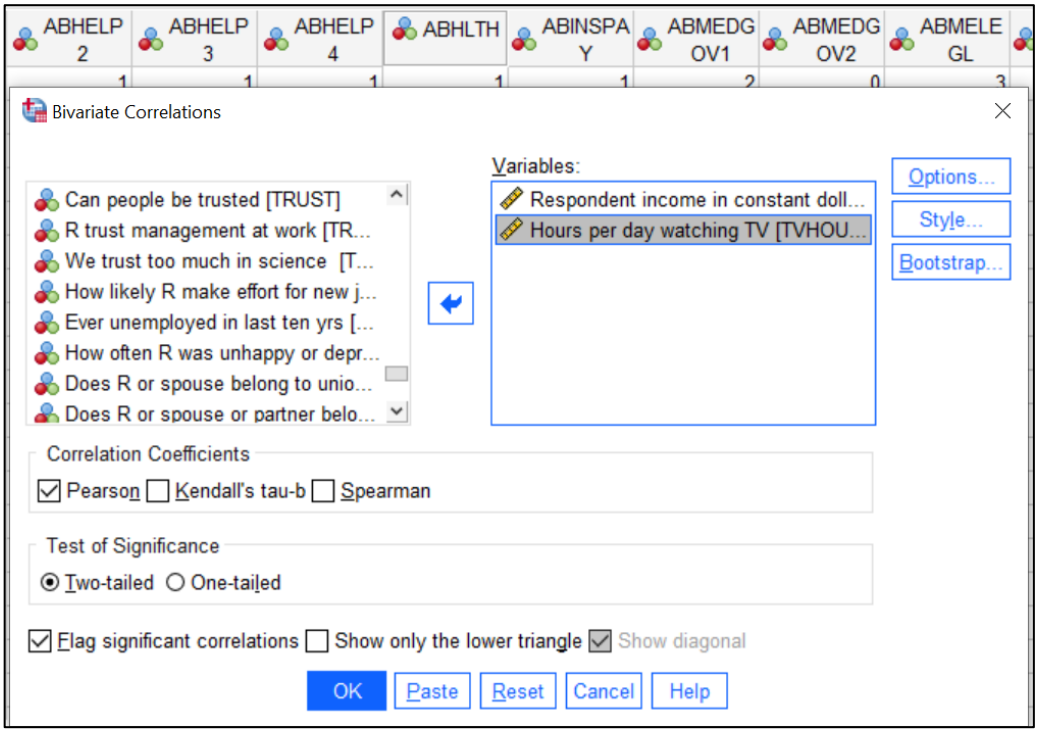
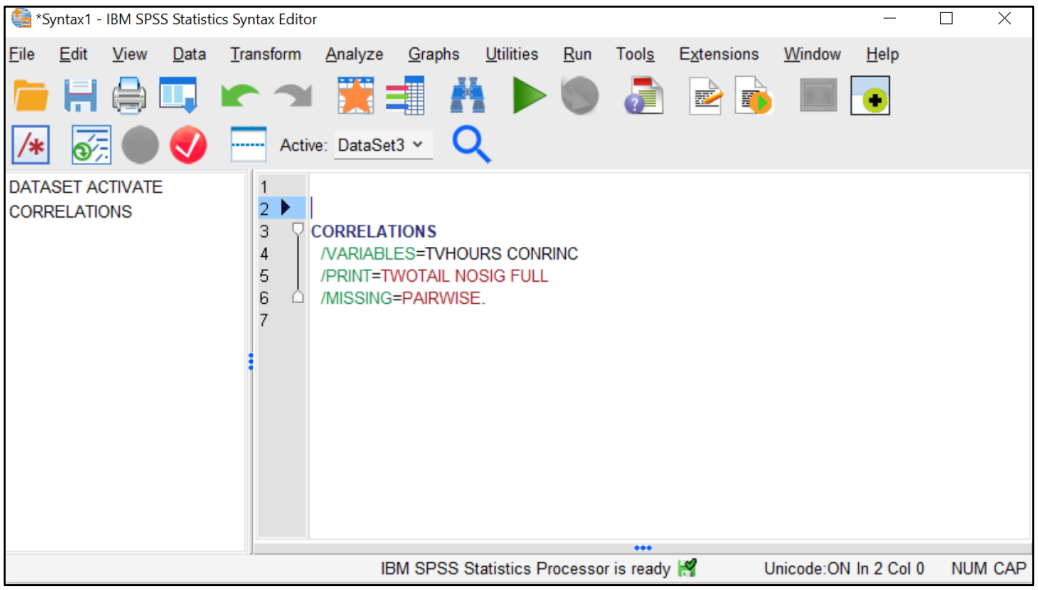


Figure 8.14. *Selecting Variables and Choosing Appropriate Correlation Coefficient.*

Alternatively, if you are dealing with a large survey that contain hundreds of various variables, it maybe easier to write a syntax and tell the program what variables and what test you would like to compute. To do that, go to **File** and open **New Syntax**. In the **Syntax Editor** type the following:



The results are given in the SPSS output. Here are the results for the first bivariate correlation:

Correlations

[DataSet3] C:\Users\mklimenko\AppData\Local\Temp\Templ_2018_spss.zip\GSS2018.sav

Correlations

		Hours per day watching TV	Respondent income in constant dollars
Hours per day watching TV	Pearson Correlation	1	-.147**
	Sig. (2-tailed)		.000
	N	1555	908
Respondent income in constant dollars	Pearson Correlation	-.147**	1
	Sig. (2-tailed)	.000	
	N	908	1363

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 8.15. *Results of Correlation #1*

The correlation between the two variables is significant (the two stars flags the statistical significance at the p value of .01 or less), and negative (see the **negative** sign).

We can also graph the values of both variables to see the pattern of this relationship. Here are the steps: select **Graphs→Legacy Dialogs→** select **Scatter/Dot**.

Choose which variable will be plotted on Y and X axis. But for logical reasons, you may want to move your predictor to be plotted on the X axis and your criterion/outcome variable on the Y axis (see Figure 8.16).

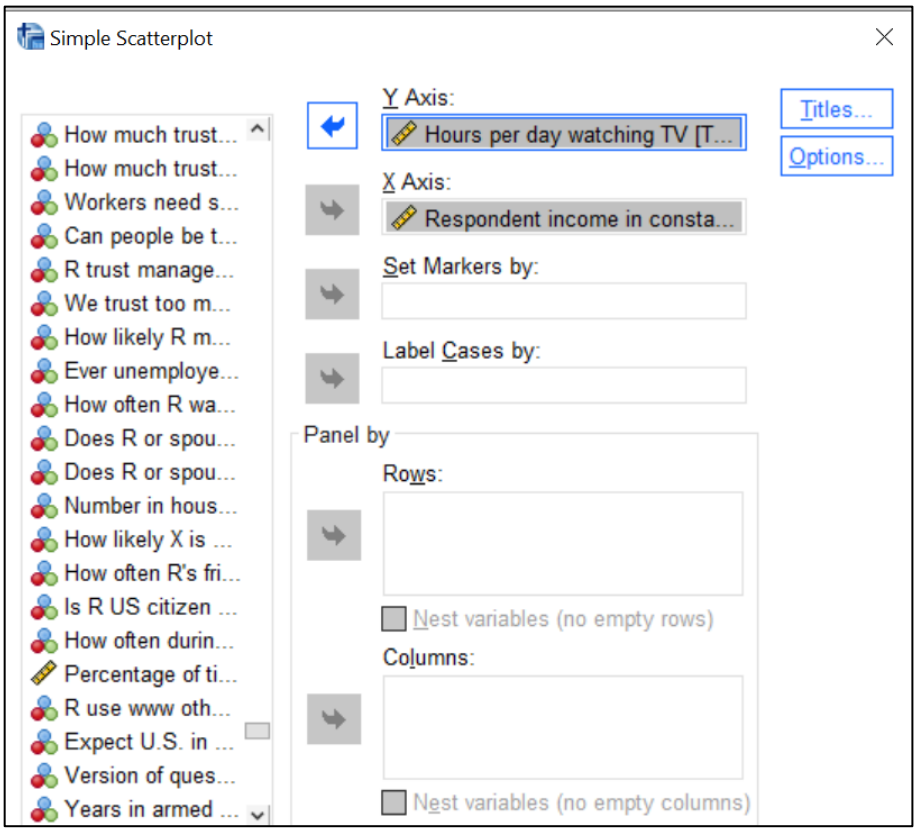


Figure 8.16. *To Graph a Scatter Plot.*

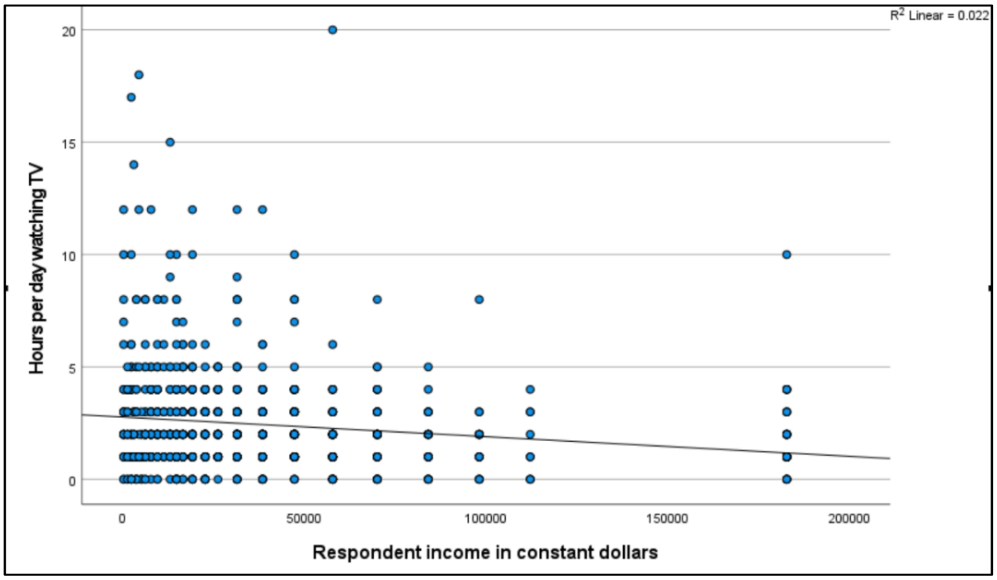


Figure 8.17. *The Scatter Plot with the Best Fitted Line (Negative Slope) to Show Negative Correlation between Respondent's Income and the Hours Per Day TV Watching.*

The negative correlation reveals that people with less income tend to watch more TV per day. Remember, we cannot make any causal inferences!

Now, let's correlate two more variables, physical level of respondent's attractiveness and his/her socio-economic status. Sounds strange? Well, some studies have found that people are perceived as more attractive when they earn more money. So let's see if we find a similar pattern. Follow the same steps: **Analyze→Correlate→** select **Bivariate**

The result of the test is shown below (see Figure 8.18).

Correlations

		R's socioeconomic index (2010)	R physical attractiveness rating
R's socioeconomic index (2010)	Pearson Correlation	1	.078**
	Sig. (2-tailed)		.001
	N	2248	1956
R physical attractiveness rating	Pearson Correlation	.078**	1
	Sig. (2-tailed)	.001	
	N	1956	2042

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 8.18. *Results of Computing Correlation #2.*

The scatter plot shows that, although not perfectly defined by the line going up, the pattern of the association between physical attractiveness and socio-economic status is positive, which is as the index of socio-economic status increases, so does the perception of respondent's physical attractiveness or as the physical attractiveness of increases so does his/her socio-economic status. Strange, of course! In sum, the two variables are positively correlated (see Figure 8.19).

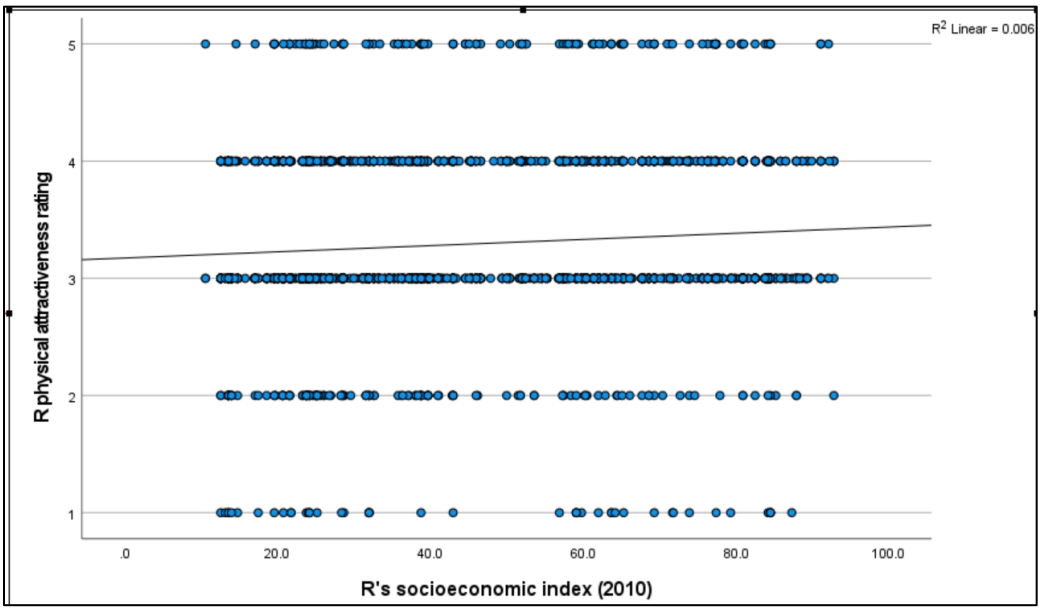
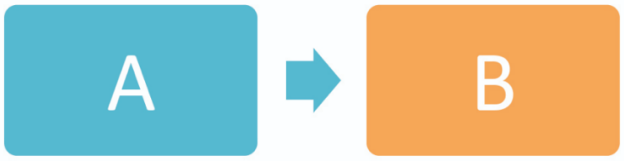


Figure 8.19. *The Scatter Plot with the Best Fitted Line (Positive Slope) to Show a Positive Correlation between Socio-Economic Status and Physical Attractiveness*

Interpreting Correlations

As was mentioned earlier, a correlation does not mean that there is a cause and effect relation. And if that is so, how should we interpret a statistically significant correlation between two variables? Several possible interpretations can be made:

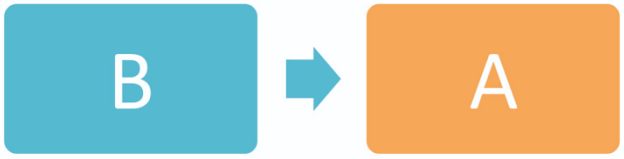
- Both variables are associated because one of them may be the cause (or one of several causes) of the other variable. For example, it is reasonable to assume that a significant correlation between exercise (A) and weight (B) found in a study may be explained by the causal effect of physical exercise on people's weight. People who exercise maintain healthy weight as a result of their physical activity. We can represent this relation as follows:



- However, if this correlation is obtained through a survey, such an interpretation cannot be confirmed and has to be treated only as a speculation.

- One can also speculate that, perhaps, people who are healthy and have lower body weight feel more comfortable and more driven to exercise. Thus, the causal factor in this

interpretation is that body weight causes people to exercise more or less. This relation can be represented as following:



- Again, if the data came from non-experimental research, this would have to remain only as speculation.

Important: The direction of causality cannot be determined from non-experimentally derived data; and the interpretation can go in both directions (AB or BA).

- It is also likely that the influence of body weight and exercise is ***bidirectional***, i.e., people who exercise lose weight and when people lose weight they continue to exercise.
- If, on the other hand, we conducted an experiment where participants in an experimental condition lost weight due to exercise, the positive correlation between exercise and weight loss would provide strong evidence that exercise caused a loss of body weight.
- In non-experimental research, there is a third possible interpretation of a correlation. The two variables can be correlated because both are caused by a third variable not included in the data or analysis. For example, suppose that the data regarding exercise and body weight was collected through a survey of University of Florida faculty members. One of the characteristics of this sample is that they, on average, have a higher income than non-faculty residing in the same local area. A third alternative explanation of the obtained correlation between exercise and body weight could be that people of higher income can afford to buy healthier food and also exercise regularly. So, it is what they eat rather than the fact that they exercise that explains their body weight. This third interpretation can be represented as following:



Of course, there is plenty of prior evidence to believe that exercising can and does reduce body weight. So, both eating healthy and exercising are related and likely to reduce body weight to some extent. However, it is possible to find two variables, that are logically or theoretically unrelated; they may still be statistically correlated because a third variable exerts its influence on both variables. We call such correlations ***spurious***. For example, apparently there is a positive correlation between the amount of ice-cream sold and the number of people who drown—a crazy correlation indeed (Johnson & Christensen, 2013, p. 395)! One explanation is that a third variable, hot weather, contributes to both instances. Otherwise, ice-cream and drowning are completely unrelated.

References

Johnson R. B., & Christensen, L. (2013). Educational research: Quantitative, qualitative, and mixed approaches(5th ed.). Sage Publications.

Smith, Tom W, Peter Marsden, Michael Hout, and Jibum Kim. General Social Surveys, 1972-2012[machine-readable data file] /Principal Investigator, Tom W. Smith; Co-Principal Investigator, Peter V. Marsden; Co-Principal Investigator, Michael Hout; Sponsored by National Science Foundation. —NORC ed.—Chicago: National Opinion Research Center [producer]; Storrs, CT: The Roper Center for Public Opinion Research, University of Connecticut [distributor], 2013.

Chapter 9: Testing a Statistical Hypothesis with SPSS Like a Scientist

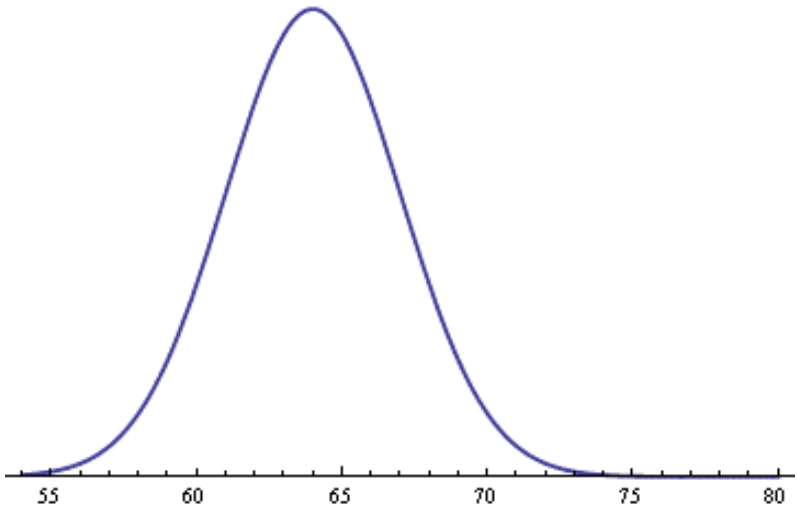
Introduction

In this chapter, you will learn how to select the appropriate inferential statistic to test your hypothesis. How to enter data into SPSS, one of the most commonly used statistical programs in psychology, will also be explained. The choice of your statistical analysis will be mainly based on the ***nature of your data*** and the ***type of your research question***.

9.1 Selection of the Right Statistical Test

To properly select the inferential analysis to test a hypothesis, one has to consider the following factors: the measurement scale of the variables, the nature of the research question, the number of the dependent and independent variables and the number of the levels of the independent variables. As you can see, the process is not the most straightforward, but I will try to simplify it as much as possible without going into details of statistics. The goal is to give you enough information to enable you to choose the right test and to compute it with the statistical software, SPSS.

First, we can broadly categorize the tests into parametric and non-parametric. All parametric tests are based on the assumption that the data is normally distributed, i.e., has a "bell-shaped curve". In other words, that the characteristic that is measured using a sample comes from the population with the normal distribution. A good example is people's height. If you sample people's height you will get a normally distributed data: most people will fall somewhere in the middle, with only few extremes to the left and to the right (see example below).



In addition, parametric tests deal with data ratio or interval data; and the variance must be homogeneous (i.e., the variance of two or more groups in the sample are equally spread). Most common parametric tests are t-Test, ANOVA, and linear regression.

Tests that fall within a ***non-parametric*** category are those that deal with data that is not normally distributed and either are nominal or ordinal. The most common non-parametric test is the chi-square test. If both of your variables are nominal or ordinal variables, you don't have to consider any other factors in order to choose the test: it must be a non-parametric test, such as the chi-square test.

As a side note, dealing with ordinal variables is a little more complicated. Researchers may purposefully try to use a scale that is either interval or nominal to avoid statistical complications that comes with ordinal data.

Next, once you know the category of the tests that you have to choose from (parametric or non-parametric), you may have to consider the nature of the research question. Most research questions fall within one of three categories: ***descriptive***, ***group comparison (differences)***, and ***relationship (association) strength***. In some cases, the same research question can be viewed as group comparison and relationship, in which case more than one statistical analysis can be used, and both should give about the same result. If the data is nominal or ordinal, the nature of the research question is less relevant as the same non-parametric test will be applied to address both group comparison and association questions as in the case of a chi-square test.

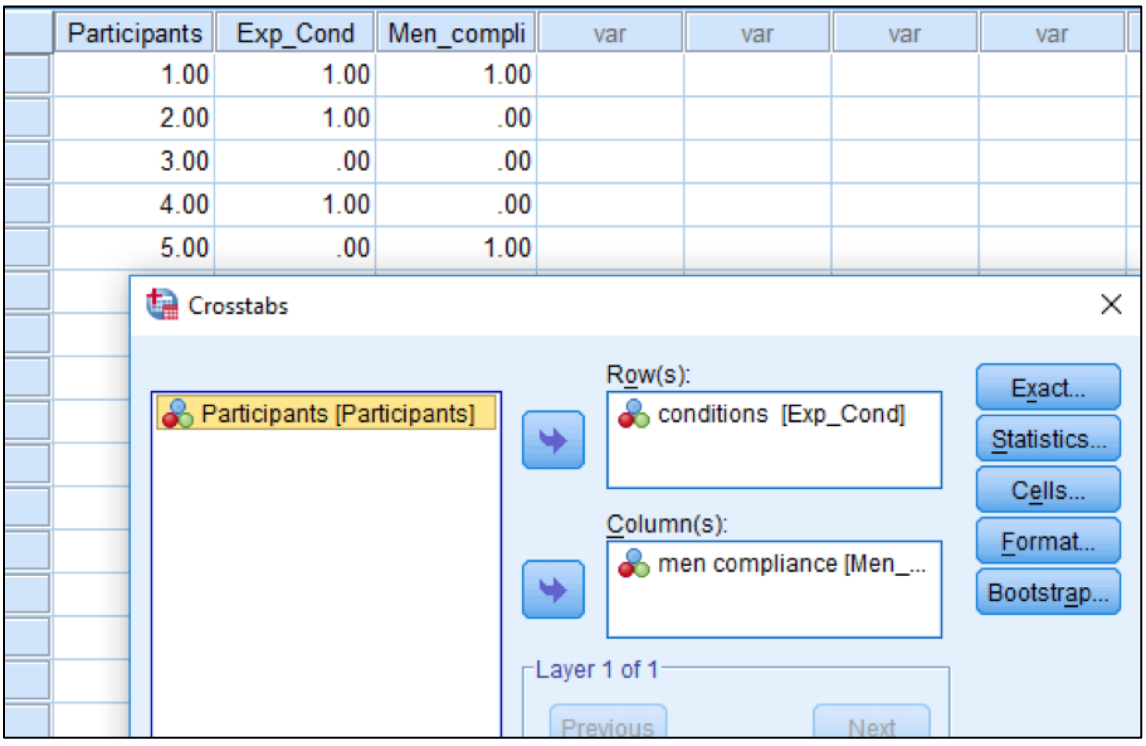
The chapter is divided by sections of parametric test of group comparison (differences), non-parametric tests of group comparison and association (the same test can test both), and parametric test of questions of associations. This should help you choose the right section and to learn the steps of conducting the test with the SPSS.

Descriptive Research Questions

Oftentimes, an uncommon or previously unknown phenomenon will have to be described before researchers can explain it. A clinical psychologist who has a patient with a rare psychological disorder will probably have to start by trying to find out more about the symptoms before he/she can theorize about the causes of the disorder and decide on the treatment. A developmental psychologist, who studies languages and wants to understand how some children acquire more advanced language skills earlier than others, will first have to describe the typical range of language skills at those particular age groups. Depending on how big, complex, or unusual the phenomenon in question is, describing it may be the primary goal of the research. But even if describing the event isn't the primary goal, most published studies usually have to describe important characteristics of the sample and other key features of the data before they proceed with hypothesis testing. Descriptive research questions will require descriptive statistics. This topic was covered in the previous chapter (refer back to Chapter 8 for details about how to compute

descriptive statistics). Similar to the inferential statistical tests, limitations of the descriptive statistics will stem from the scale of the variables. As was already noted, nominal data is the most limited in the kinds of descriptive statistics a researcher can use to describe the sample. Recall that a nominal variable simply categorizes information into categories or groups (e.g., 'agree' and 'disagree,' 'female' and 'male,' or 'young' and 'old'). These categories will typically be coded with numbers, such as '1' for agree and '0' for 'disagree.' But the numbers are not meant to represent any meaningful numeric information other than to distinguish the categories. Most typical descriptive statistics that can be computed with nominal variables is the mode and the cross-tabulation. The mode tells you what category is most frequent in a given sample (e.g., males or females). The cross-tabulation represents a joint frequency; for example, if data has two nominal variables, gender (male and female) and opinion (agree and disagree), a cross-tabulation of gender and opinion will tell you how many females or males had the 'agree' and 'disagree' opinion. To compute a cross tabulation in SPSS, go to Analyze→Descriptive Statistics→Crosstabs (see below).

In my example below, I have two nominal variables, *Exp_Cond* represents two conditions, '1' experimental and '0' control, and *Men_compli* that represents a nominal dependent variable, which is men who complied with a request (coded as '1'), and men who did not comply (coded as '0'). Move one variable to the Row(s) and the other one to the Column(s). Finally, click on OK.



The output will give you the following crosstabulation table:

conditions * men compliance Crosstabulation

Count

		men compliance		
		did not comply	complied	Total
conditions	control	1	1	2
	experimental	2	1	3
Total		3	2	5

The crosstabulation table shows that two out of three men in the experimental condition did not comply and one did; while one man complied and one did not comply in the control condition. Remember that this only gives you a descriptive statistic, and this information does not permit making any inferences about the correctness of the hypothesis. In other words, even though more men did not comply in the experimental condition (supposedly this is what was predicted), this does not mean that this didn't happen by chance alone, we need to know how small the probability of obtaining this result by chance. The inferential tests, which we will discuss next, will compute this statistic for us. Also, note that the sample must be a lot bigger than in my example to make any inferences.

As you can see next, SPSS can, or in some tests, will compute automatically the descriptive along with the inferential statistics, to help with the interpretation of the results.

For more information on descriptive statistics, go back to chapter 3. We now move on to the selection and computation of the ***parametric*** inferential statistics.

9.2 Research Questions of Group Comparison (Differences)

Are there any differences in how the female and male brains process emotional information? Do people who play violent video games tend to express more aggression than those who don't? These are just a couple of examples of *group comparison* types of research questions, i.e., comparing groups on some selected measure such as, aggression, brain processing, etc. Group comparison questions are usually associated with experimental studies, in which researchers can randomly assign people to one, two, or more groups and compare their performance on some chosen measure. However, ***non-experimental*** *studies* can also ask research questions about differences between groups, even if they don't experimentally manipulate their independent variable(s). A researcher might conduct a survey to ask people how frequently they play violent video games and

how often or how intensely they become angry while playing them. You can then chart the frequency or the intensity of anger expressed by those who play and who don't play violent video games. The key here is the comparison of two or several groups on some measure of independent or criterion variable(s).

The most commonly applied statistical analyses that test group differences are the ***t* test** and **ANOVA**. The choice will depend on the number of independent/predictor variables, the number of levels each independent/predictor variable has, and the measurement scale of the dependent/criterion variable. Additionally, you will need to consider whether the design of the experiment, if it is an *experiment*, or the nature of your research question, if it is a correlational study, is between, or within, subject design.

Choosing appropriate inferential statistic to test group differences: The choice of *inferential statistic* to be used will be based on the factors of, the number of independent variables and the number of levels that the independent variable has, the scale of the dependent variable, and whether the design is between, or within, subjects.

Table 1 illustrates the summary of these considerations and identifies appropriate statistical analyses after considering the above factors.

*Table 1:*Inferential Statistical Analyses for Comparison Group Questions

Between-Subjects Design: Independent samples			Within-Subjects Design: Repeated measures or related samples
Interval or Ratio (Continuous) Dependent Variable: PARAMETRIC STATISTICS			
One independent variable with 2 levels (groups)	One independent variable with 2 or more levels (groups)	Two independent variables	
Independent Samples *t* Test	**One-Way ANOVA (*F*)**	**Two-Way Factorial ANOVA (3*F*s)**	**Paired Samples *t* Test**
Nominal or Ordinal Dependent Variable: NON-PARAMETRIC STATISTICS			
Chi-square	**Chi-square**		

**This table was adapted from Morgan, Griego, & Gloeckner (2001).*

In the next section the process of computing each ***parametric*** test identified in Table 1 is explained. Keep in mind that although these are the most commonly used statistical procedures, they are not the only procedures available to researchers. Finally, since the focus of this book is to offer a practical and quick guide to research methods in psychology, the mathematical computations involved in computing these statistics will not be covered. Instead, because the logic and purpose behind each statistical analysis are fairly easily understood, the steps of computing these mathematical computations in SPSS is an easy process.

Important information to keep in mind: SPSS uses the terms, 'dependent' and 'independent' variable, irrespective of whether the data came from an experimental or correlational study. However, this distinction needs to be clearly made when you report your results in a research paper. If your data was derived from a correlational study, use the terms *predictor* and *criterion* variable when describing your methods and statistical analyses.

Some General Guidelines for Reporting Statistical Results

- Report numbers between 10 and 100 by rounding to one decimal place (e.g., 12.2)
- Report numbers between 0.10 and 10 by rounding to two decimal places (e.g., 1.22)
- Report numbers that are less than 0.10 to three or more (to have non-zero number) decimals (e.g., 0.0000012)
- Report the exact *p*-value, omitting the first zero. If the *p*-value is .058, then report, p =.058
- Most common statistics are reported as *statistic* (degrees of freedom)=value, *p*=value, *effect size statistic*=value

9.3 Parametric Tests for Questions of Group Differences

Independent Samples t Test (Between Groups)

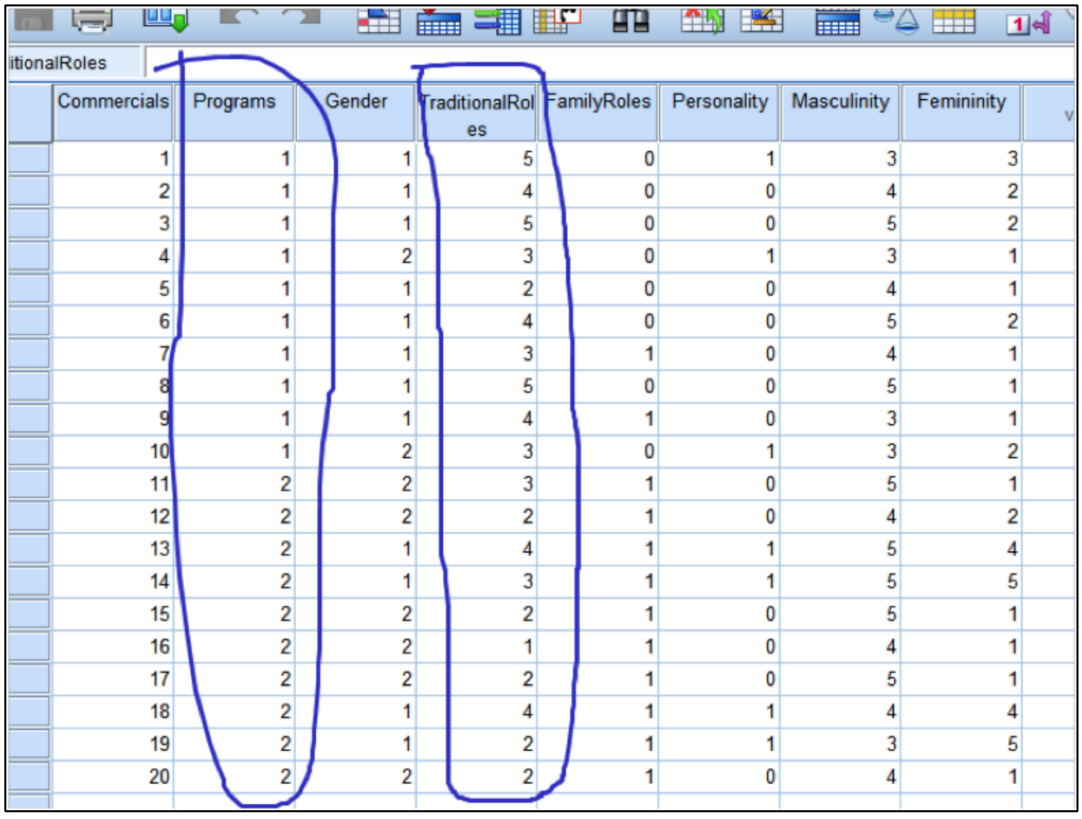
Conduct an independent samples *t* test if you have a normally distributed, continuous dependent variable (interval or ratio scale) and one independent variable with two levels or groups (e.g., males and females). Let's use a hypothetical data set to show the logic and the computations of this test with SPSS. Suppose a researcher wants to know the following:

Example 1: *Do TV commercials differ in terms of how stereotypically traditional their main characters are portrayed when aired during* ***political*** *and* ***entertainment*** *programs?*

This is a ***between groups*** comparison research question because it seeks to compare TV commercials aired during entertainment and political programs. The predictor variable is the

programming type, because there is evidence that TV commercials will vary in terms of how stereotypically traditional their characters are portrayed based on the audience watching certain types of programs. Perhaps people who tend to watch political programs are older and/or more conservative than the audience of the entertainment programs, and so they will respond more positively to TV commercials portraying characters in more traditional roles. Notice that the hypothesis is *two-tailed*, i.e., not specifying which group will have a higher mean. And even though the mean for the political group should possibly be higher, it will be informative to see if it goes in the opposite direction. The predictor variable has two levels (groups): political and entertainment, and my dependent variable, traditional role portrayal, is a scale variable (see data set TV commercials below). Notice also that there is no interest in differentiating between male and female characters. The research has hypothesized that regardless of the characters' gender, they will be portrayed either less or more traditional, and, that the kinds of programs (proxy for the type of audience) for whom the TV commercials are intended is what drives the portrayal of the characters as more or less traditional.

Here is what the dataset (TVcommercials.sav) looks like in the SPSS file:



itionalRoles

	Commercials	Programs	Gender	TraditionalRoles	FamilyRoles	Personality	Masculinity	Femininity
1	1	1	1	5	0	1	3	3
2	2	1	1	4	0	0	4	2
3	3	1	1	5	0	0	5	2
4	4	1	2	3	0	1	3	1
5	5	1	1	2	0	0	4	1
6	6	1	1	4	0	0	5	2
7	7	1	1	3	1	0	4	1
8	8	1	1	5	0	0	5	1
9	9	1	1	4	1	0	3	1
10	10	1	2	3	0	1	3	2
11	11	2	2	3	1	0	5	1
12	12	2	2	2	1	0	4	2
13	13	2	1	4	1	1	5	4
14	14	2	1	3	1	1	5	5
15	15	2	2	2	1	0	5	1
16	16	2	2	1	1	0	4	1
17	17	2	2	2	1	0	5	1
18	18	2	1	4	1	1	4	4
19	19	2	1	2	1	1	3	5
20	20	2	2	2	1	0	4	1

Figure 1. *SPSS Variable View for TV commercials Data Set.*

I circled both, the predictor and criterion variables – the only two variables we are interested in this example. You may also notice the difference in the measures of scale between the two variables: the criterion variable, i.e., the traditional role, is a normally distributed, continuous variable ranging from '1' (the least traditional role portrayal) to '5' (the most traditional role portrayal). The predictor variable, the program types, is a nominal variable, coded as '1' for political and '2' for entertainment. It is crucial that, before you proceed with your analysis, you understand the nature of your variables and what their values represent. The research question and the nature of the variables let me know that the appropriate analysis is to conduct an independent samples *t* Test.

To run an Independent Sample *t* Test (Between Groups Design), complete the following:

- Click on Analyze→Compare means→Independent Samples T Test.
- Move criterion variable 'Traditional Roles' to the Test variables window and the predictor variable 'Programs' to the Grouping variable window (see Figure 2)

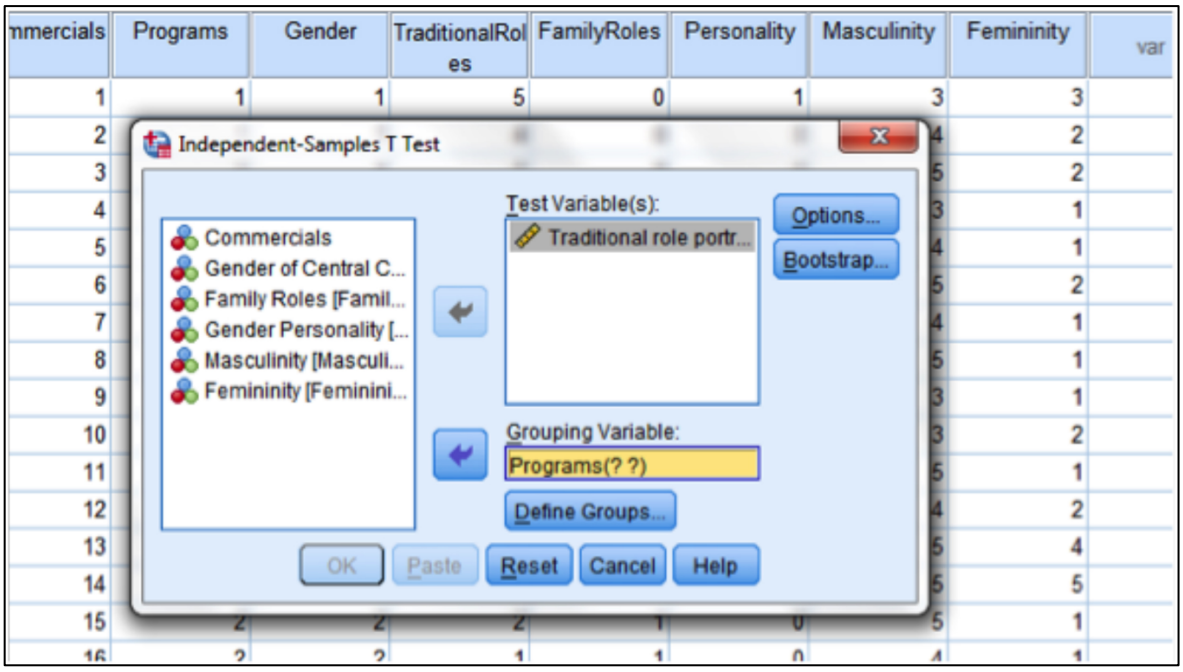


Figure 2. *Independent-Samples T Test Variables Box.*

Next, the groups, or, levels, are defined as political and entertainment by clicking on Define Groups (see Figure 3).

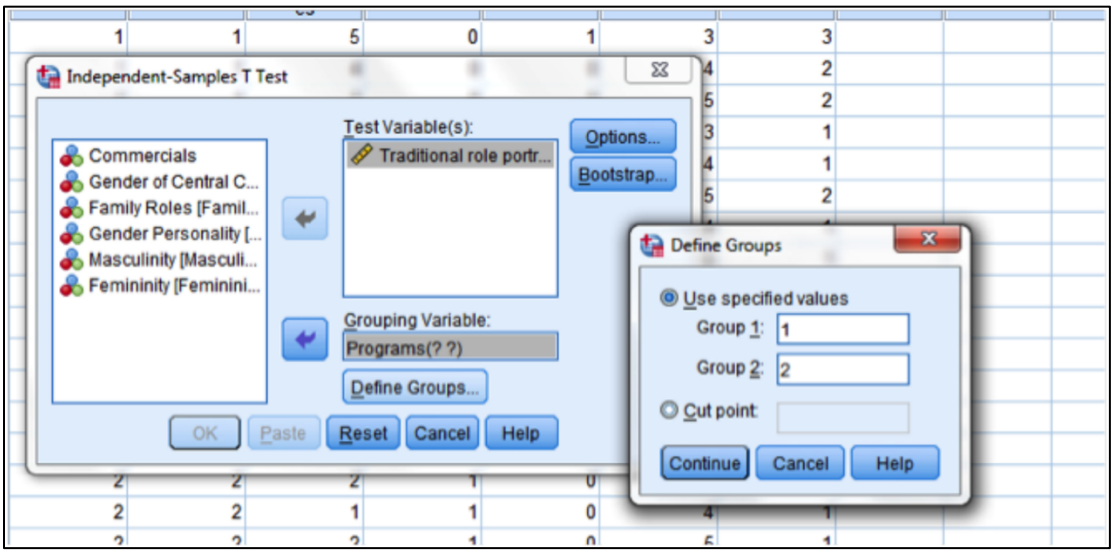


Figure 3 .*Independent-Samples T Test: Defining Variables.*

If I click continue→OK, the program will immediately calculate the results. However, I can also click on Paste and the program will proffer the syntax which saves the steps of my analysis (see Figure 4). This option is very useful when a researcher runs multiple, different, statistical analyses or manipulates the data because it saves all of your steps and manipulations. This option can also quickly recreate your analyses or the manipulations and recalculate the results. To do so, simply reopen your syntax and click on the green arrow button on the top, and the program will recreate everything.

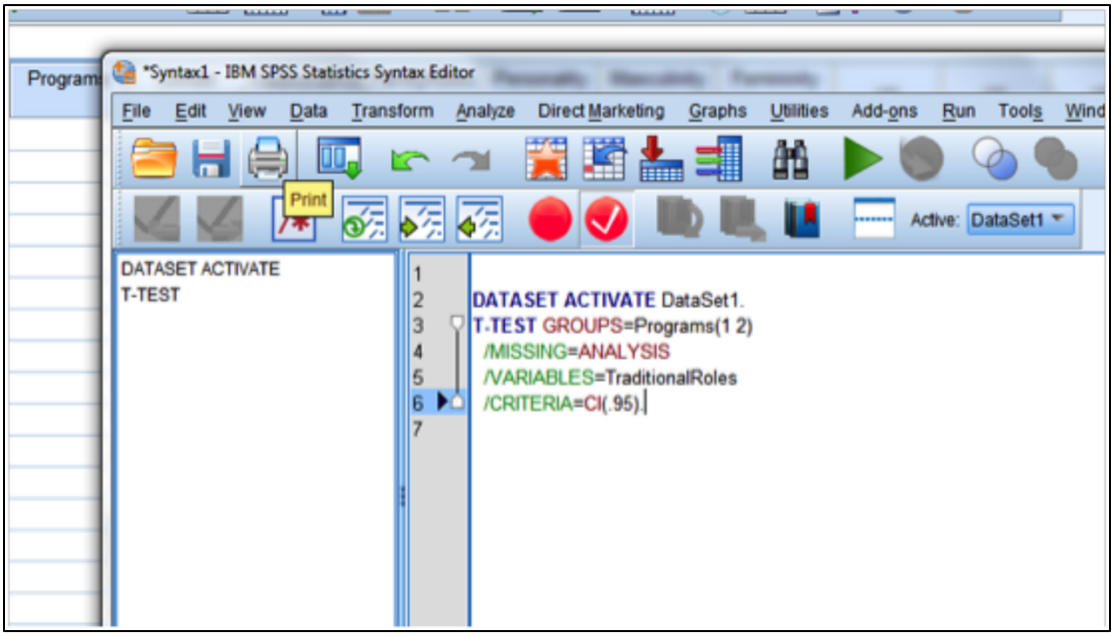


Figure 4. *Syntax for Independent-Samples T Test.*

Let's now turn our attention to the results of the analysis. Here is what the 'output' of my results looks like (see Tables 2 and 3). The first table in the SPSS output (Table 2) gives the descriptive statistics. It tells me that I have two groups, or levels: political and entertainment programs. This table also shows that I have an equal number of commercials in each group, ten commercials in each. Additionally, it shows that their means are slightly different; that the mean of traditional role portrayal is 3.80 and 2.50 for TV commercials aired during the political and the entertainment programs, respectively. So what do these averages represent? To be able to make sense of these descriptive statistics, one has to remember the way my variables have been coded. Recall that my codes for traditional role portrayal range from, '1', being the least traditional, to '5', the most traditional portrayal. So, the higher mean of traditional role portrayal in political programs means that the characters (men or women) were portrayed more traditionally during political than entertainment programs.

Table 2. SPSS Output of Descriptive Statistics

Group Statistics

	Program Types	N	Mean	Std. Deviation	Std. Error Mean
Traditional role portrayal	Political	10	3.80	1.033	.327
	Entertainment	10	2.50	.972	.307

Depending on the nature of the data, additional descriptive statistics may also be useful, and in some instances, even more informative than simply looking at the means. It may be also informative to find the frequency distribution of the five codes of traditional role portrayal. To get that information, I can do the following:

- Analyze→Descriptive Statistics→Frequencies
- Move Traditional Roles to Variable(s) box
- Click on Statistics→check **range** (you can also check **mean**)
- Click on Continue and Ok Table 3 displays the results.

*Table 3.*SPSS Output of the Frequency Counts of Traditional Role Portrayal Variable

Traditional role portrayal

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1	5.0	5.0	5.0
	2	6	30.0	30.0	35.0
	3	5	25.0	25.0	60.0
	4	5	25.0	25.0	85.0
	5	3	15.0	15.0	100.0
	Total	20	100.0	100.0	

These descriptive statistics illustrate that the majority of the TV commercials in my sample fell somewhere between '2' and '4', which is presumably the middle range of my coding scheme for the traditional role. In other words, the characters seem to be moderately traditional in their portrayal.

However, the results of my descriptive statistics do not permit me to answer my main research question just yet. I will need to look at the results of my inferential statistic. Let's state our hypothesis in statistical terms by specifying the null and the alternative hypotheses:

H_0: No difference exists between the two means (traditional role of TV commercials aired during political and entertainment programs). Programs have no effect on whether TV characters are portrayed stereotypically traditional.

H_1: The difference between the two means (traditional role of TV commercials aired during political and entertainment programs) is significant; programs have an effect on whether TV characters are portrayed stereotypically traditional.

Recall that to in order to find support to my hypothesis I need to reject the null hypothesis. The t test will determine whether any observed differences between the two means are big enough to reject the null hypothesis and to conclude that the observed difference is, in fact, due to the effect of the programs, and not simply to chance.

Statistical Significance

Let's look at the second, our main table, in the SPSS output (see Table 4) which gives us the results of the two tests: the Levene test and the actual *t* test that measures my research hypothesis. The Levene test is displayed in the two left columns. This tests the assumption that the variances of my two groups are equal (remember that this is one of the important assumptions of the parametric statistical analyses). The null hypothesis here is that the variances are equal. If the *F* test is not significant, I cannot reject the null hypothesis. This indicates that the assumption of equal variances is not violated (that's what we want)—the variances are indeed equal. I find that the *F* is .031 and it is not statistically significant (.862 >.05). This shows that I should look at the results that are displayed in the **equal variance assumed** line (equal variances assumption was not violated) to read the *t* test and other related statistics: *t*= 2.899, degrees of freedom (*df*) = 18, and *p* = .010. Because *p* value is less than .05, I can reject the null hypothesis and conclude that the two means are statistically significant!

*Table 4.*SPSS Table with Results of the Independent-Samples *T* Test

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Traditional role portrayal	Equal variances assumed	.031	.862	2.899	18	.010	1.300	.448	.358	2.242
	Equal variances not assumed			2.899	17.934	.010	1.300	.448	.358	2.242

Now, let's translate this statement into more meaningful language by returning to my original question: Do TV commercials differ in terms of how stereotypically traditional their main characters are portrayed when they are aired during **political** and **entertainment** programs? The answer is 'yes.' Based on my results, I can conclude that the TV commercials tend to portray their main characters in more traditional roles when they are aired during political than during entertainment programs.

Practical Significance

Although the difference between the two groups is ***statistically*** significant, it doesn't necessarily tell us about the ***practical*** significance of this finding. In other words, we still don't know how very large the difference is in the portrayal of the main characters. To answer this question, I have to measure the effect size (*d*). This has to be computed manually, by subtracting the mean of one group (i.e., the mean of entertainment group) from group two (i.e., political) and dividing by the pooled standard deviation of both, as follows:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}}$$

where

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Let's plug in the numbers:
$(S)^2_p=(1.033)^2(10) + (.972)^2(10)/(10 + 10 - 2)$
$(S)^2_p=1.06(10) + 0.94(10)/18$
$(S)_p=11.12$
$d= (3.80 - 2.50)/11.12$
$d = 0.11$

In order to interpret this number in social sciences, we use the following, Cohen's (1988), rule of thumb (here we are only looking at *d*).

Small effect: *d*=.2, *r*=.1, eta^2 of .01
Medium effect: *d*=.5, *r*=.3, eta^2 of .09
Large effect: *d*=.8, *r*=.5, eta^2 of .25

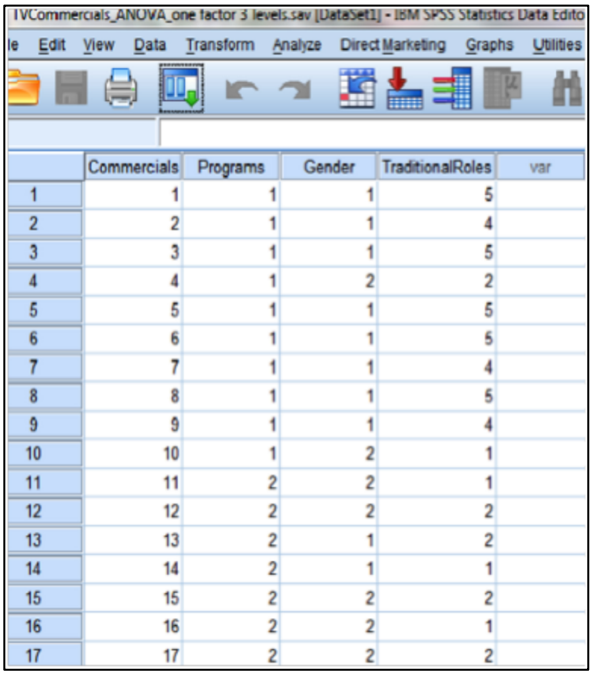
Know that what is considered a large effect size in social and behavioral sciences may not be considered large in other disciplines, for various reasons. The effect size in my sample turns out to be very small, and this may not be very surprising given the small differences between the means. This does affect the interpretation of the findings, specifically, its practical significance.

Reporting Your Results: Three important pieces of information should be reported: the t-statistic, df (degrees of freedom), and p-value. In addition, it is a good idea to report the means of both groups. I would report them as the following statements: The t test shows that the main characters in TV commercials, aired during political programs, were portrayed significantly more traditional (M = 3.80) than the characters in commercials, aired during entertainment programs (M = 2.50), $t(18) = 2.89$, $p < .05$. Finally, briefly summarize these statistical findings in a more commonly used language so that people, who are not so savvy in statistics, can understand your results. For example, I would finish my results section with the following statement: As was hypothesized, main characters in TV commercials aired during political programs were portrayed in more traditional roles [insert examples] than the main characters of TV commercials aired during entertainment programs.

One-Way (Single Factor) ANOVA

ANOVA (analysis of variance) can also be used to compare two sample means; results will be identical to a *t* Test. But you must use ANOVA if you compare more than two means. Example 2: Suppose I decided to have three groups of programs: political, reality, and other entertainment programs (see TV commercials_ANOVA_one factor 3 levels.sav data set). To perform a three-group comparison, I will have to run a One-Way (or Single Factor) ANOVA (because I only have one independent variable with three levels/groups). The logic of this analysis is similar to a *t* test, except that we are now comparing three, rather than two, groups.

Here is what this hypothetical data looks like (see Figure 5).



	Commercials	Programs	Gender	TraditionalRoles	var
1	1	1	1	5	
2	2	1	1	4	
3	3	1	1	5	
4	4	1	2	2	
5	5	1	1	5	
6	6	1	1	5	
7	7	1	1	4	
8	8	1	1	5	
9	9	1	1	4	
10	10	1	2	1	
11	11	2	2	1	
12	12	2	2	2	
13	13	2	1	2	
14	14	2	1	1	
15	15	2	2	2	
16	16	2	2	1	
17	17	2	2	2	

Figure 5. *SPSS Variable View for TV commercials_ANOVA_one factor 3 levels Data Set.*

The following are the steps to compute a One-Way ANOVA using SPSS:

- Analyze→Compare Means→One-Way ANOVA
- Move Traditional roles to the Dependent Variable box, and programs to Factor box
- Click Options→Descriptives and Homogeneity of variance[7]

Let's interpret the results. As always, start by examining your descriptive statistics first. This examination reveals that all three means are different. The difference is especially noticeable between the political and the reality programs (see Table 5). However, it doesn't tell us whether these differences are, indeed, ***statistically*** significant. We will have to look at the ANOVA table. But first, let's take a look at the Levene's test table (Table 6).

*Table 5.*SPSS Output of the Descriptive Statistics of the Main Variables

Descriptives

Traditional role portrayal

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Political	10	4.00	1.414	.447	2.99	5.01	1	5
Reality	10	1.70	.675	.213	1.22	2.18	1	3
Other Entertainment	10	3.30	1.418	.448	2.29	4.31	1	5
Total	30	3.00	1.531	.280	2.43	3.57	1	5

*Table 6.*SPSS Output of Levene's Statistic

Test of Homogeneity of Variances

Traditional role portrayal

Levene Statistic	df1	df2	Sig.
2.090	2	27	.143

[7] See my discussion on the next page about why we have to test for homogeneity of variance.

Because this test uses ***variances*** to determine if the group means are statistically different, one of the assumptions in the analysis of variance test is that the variances of the groups that are being tested or compared are equal or homogeneous. And, although small deviations in the variances are generally not a big problem in ANOVA, it is a good idea to make sure that the data meet this assumption. An SPSS can automatically verify this by performing Leven's test (one of the available hypothesis tests). And, just like any other statistical hypothesis testing, it tests the null hypothesis. However, in this test, it is what we don't want to reject—which is, that the variances are, indeed, equal. Failure to reject the null means the two variances are homogeneous (that's what we want). Let's look at the results of our Levene's test. Because the test is not statistically significant (p=.143), we do not reject the null hypothesis and can conclude that the assumption of equal variances was not violated. And since a one-way ANOVA test was appropriate, we can move to the interpretation of its results.[8]

Let's look at the results of the ANOVA table (see Table 7).

Table 7. SPSS Output of One-Way ANOVA

ANOVA

Traditional role portrayal

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	27.800	2	13.900	9.336	.001
Within Groups	40.200	27	1.489		
Total	68.000	29			

The F statistic is 9.34 and it is significant (p=0.001). This is because the differences between the three groups are not due to chance but to the effect of programs. It is important to understand that this statistic only tells us that, at least two of three groups are significantly different, and it does not tell us which groups are different from each other. The descriptive statistics do suggest that there seems to be a big difference between political and reality programs and possibly between reality and other entertainment groups. However, we would not know this unless we conduct a follow-up test. Luckily, SPSS is capable of computing several follow-up, also known as *post hoc*, tests. I will use the **Tukey**, one of the common post hoc analyses; appropriate to be used when variances are equal.

[8] If the assumption of homogeneity of variance is violated, the data may have to be transformed or tested using non-parametric analysis.

Post Hoc Test: Tukey

To compute post hoc, or, Tukey, first repeat the same steps as you did to compute a One-Way ANOVA. Unclick, Descriptives and Homogeneity options (you will not need them for this test). Click on Post Hoc and select, **Tukey**.

Table 8. SPSS Output of Post Hoc Tukey Test

Multiple Comparisons

Dependent Variable: Traditional role portrayal

Tukey HSD

(I) Program Types	(J) Program Types	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Political	Reality	2.300*	.546	.001	.95	3.65
	Other Entertainment	.700	.546	.417	-.65	2.05
Reality	Political	-2.300*	.546	.001	-3.65	-.95
	Other Entertainment	-1.600*	.546	.018	-2.95	-.25
Other Entertainment	Political	-.700	.546	.417	-2.05	.65
	Reality	1.600*	.546	.018	.25	2.95

*. The mean difference is significant at the 0.05 level.

The tests compute the differences between the three means (some of them are duplicated). I circled the ones that we should be looking at (see Table 8): *political* and *reality*, *political*, and *other entertainment*, and *reality* and *other entertainment*. All but the difference between political and other entertainment programs is not significantly different (p=.417). This means that significant differences exist between how traditionally TV commercials' characters are portrayed. As we expected, political programs tend to have TV commercials where characters are more traditional than the characters in commercials aired during entertainment programs. But no such difference exists between commercials aired during reality programs and other entertainment programs.

Reporting Your Results:

The following statistics have to be reported when you write your results: the F-statistic, df (degrees of freedom for the between- and within-groups), and p-value. These statistics can be presented as following: A one-way ANOVA was conducted to test differences between TV commercials aired during political, reality and other entertainment programs in terms of how stereotypically traditional their characters are portrayed. A significant effect of programs on characters' portrayal in traditional roles was found, F (2, 27) = 9.34, p<.001. i="">Since the overall test shows that the groups differ in the traditional portrayal of the characters in the commercials, you will also need to report the results of the post hoc test to further explain which groups were significantly different.

Post hoc analysis using the Tukey test revealed that the mean score for reality programs (M=1.70, SD=.68) was significantly different from political ones (M=4, SD=1.41), and also from other entertainment programs (M=3.30, SD=1.42), but the political mean score was not significantly different from other entertainment programs. The information about the means (M) and the standard deviations (SD) can be found in Table 5, and the significance tests of comparison are found in Table 8.

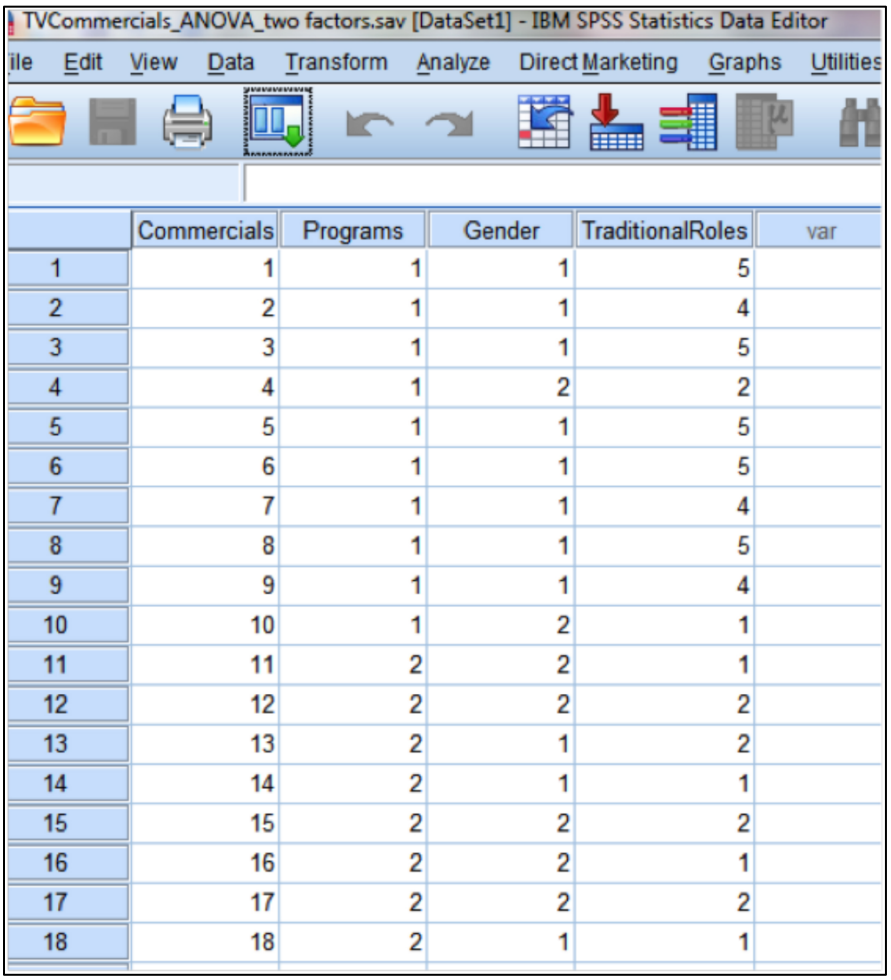
Finally, finish your results section by explaining, in common language, what all these findings suggest. For example, I might include the following statement: These findings support the hypothesis that the TV commercials' main characters are portrayed differently when they are in commercials aired during political, reality, and other types of entertainment programs. Specifically, characters seem to be the least traditionally portrayed in commercials aired during reality shows. However, interestingly enough, no such differences were found between characters in commercials aired during other entertainment or political programs. The fact that the characters' portrayal in commercials during reality shows is different from other shows suggest that the audiences of reality shows may be the least traditional. Any further speculations should be left for the discussion section of your research paper.

Factorial (2-Way) ANOVA

Suppose that in addition to the types of TV programs for which commercials are usually aired, the traditional portrayal of characters is hypothesized to be dependent on their gender. So the research question can be restated as following:

Example 3: Do TV commercials differ in terms of how stereotypically traditional their main characters are portrayed when the commercials are aired during **political**, **reality** or **other entertainment** programs, and is the gender of the character a consideration?

This question requires adding a second predictor variable (factor)—gender. Thus, I now have two predictor variables and one (interval) dependent variable. I will use a fictitious data set, TV Commercials_ANOVA_two factors, to illustrate the testing of this type of data, (see Figure 6).



	Commercials	Programs	Gender	TraditionalRoles	var
1	1	1	1	5	
2	2	1	1	4	
3	3	1	1	5	
4	4	1	2	2	
5	5	1	1	5	
6	6	1	1	5	
7	7	1	1	4	
8	8	1	1	5	
9	9	1	1	4	
10	10	1	2	1	
11	11	2	2	1	
12	12	2	2	2	
13	13	2	1	2	
14	14	2	1	1	
15	15	2	2	2	
16	16	2	2	1	
17	17	2	2	2	
18	18	2	1	1	

Figure 6. *SPSS Variable View for TV Commercials_ANOVA_two factors Data Set.*

Notice that my second predictor variable, *gender*, is also a nominal variable, where '1' stands for female characters and '2' for male characters. Everything else is the same as in the previous example.[9] My research goal is to find out if both predictors significantly affect how men and women are portrayed in TV commercials. Another way to call this design is a 2 x 3 factor design, where:

[9] Some values have been changed slightly.

- Factor 1: Gender
    - Male
    - Female
- Factor 2: Program Types
    - Political
    - Reality
    - Other entertainment

The following are the steps to run this analysis:

- Analyze→General Linear Model→Univariate
- I move Traditional roles to the Dependent variable box, and the two predictor variables, Gender and Programs to the Fixed Factor(s) box (see Figure 7)

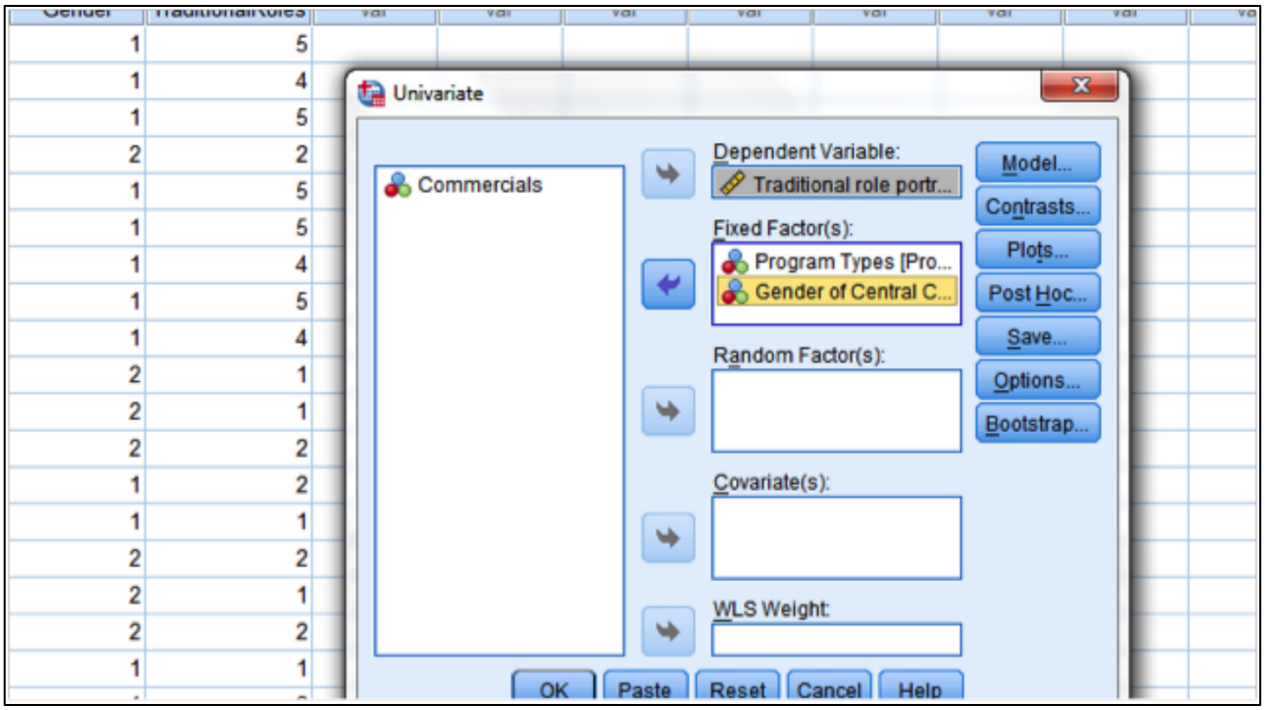


Figure 7. *SPSS General Linear Model: Two-Factor ANOVA.*

I will also need to specify more options:

- Click on Plots, move Program to Horizontal axis[10] and Gender to the Separate Lines (the purpose of this plot will be clear once you see what it displays), press Add and Continue.
- I also click on Options→click on Descriptive statistics and Estimates of effect size. Notice also, that this analysis allows us to automatically, and very easily, compute the effect size.

The first table in the output gives us the frequency counts for the two predictor variables (see Table 9). The next table provides the means and the standard deviations, cross-tabulated by gender and program type (see Table 10). This cross-tabulation table is the key to answering our research question. We will come back to it when we evaluate the plot profile. But for now, examination of the cross-tabulation table reveals the following patterns: Female characters in political and other entertainment categories have the highest means on the measure of traditional role portrayal, while male characters have the higher mean on the same measure when compared with female characters in the reality program category (see Table 10).

Table 9. SPSS Frequency Counts of the Independent Variables

Between-Subjects Factors

		Value Label	N
Gender of Central Character	1	female	16
	2	male	14
Program Types	1	Political	10
	2	Reality	10
	3	Other Entertainment	10

Table 10. *SPSS Descriptive Statistics Cross-Tabulated by Gender and Program Types*

[10] It is easier to read the plot if a categorical variable is placed on the horizontal axis. In this example, both variables are categorical, so it does not make much difference.

Descriptive Statistics

Dependent Variable: Traditional role portrayal

Gender of Central Character	Program Types	Mean	Std. Deviation	N
female	Political	4.63	.518	8
	Reality	1.50	.577	4
	Other Entertainment	4.00	.000	4
	Total	3.69	1.401	16
male	Political	1.50	.707	2
	Reality	1.83	.753	6
	Other Entertainment	2.83	1.722	6
	Total	2.21	1.311	14
Total	Political	4.00	1.414	10
	Reality	1.70	.675	10
	Other Entertainment	3.30	1.418	10
	Total	3.00	1.531	30

Now, let's see if these interesting patterns are actually ***statistically*** significant. This information is displayed in the table, *Tests of Between-Subjects Effects* (see Table 11) . Notice, that in addition to my two predictor variables, ***program types*** and ***gender***, one more variable is listed—***Programs*Gender***. Program and Gender variables represent the main (and independent) effects of program type and gender, respectively. In other words, it tells us whether program type has an effect on the traditional portrayal of the characters in TV commercials without considering the gender of the characters. Similarly, the main effect of gender tells us whether being female or male makes a difference in terms of how traditionally the characters are portrayed regardless of the programming. Both main effects are significant (see the p- values for both factors in Table 11).

However, before I can make any conclusions about the main effects of program types and gender, I have to consider the interaction effect between the two predictor variables—Programs*Gender. This is exactly what my main research question was—to see if both gender AND program type would have an effect on how traditionally characters are portrayed in commercials.

***Important: If the interaction effect is significant, then the main effects should not be interpreted. The results should be discussed only in terms of the significance of the interaction between the two predictor variables.**

Let's take a look at the key table in the SPSS output that displays the main results of the two-factor ANOVA, Table 11. Notice that we use the F values (not t-value) and the F distribution in this test. All three *F*s are statistically significant. But, the interaction effect of program type by gender is also significant. This means that I should interpret the interaction effect only.

*Table 11.*SPSS Results Output of Two-Factor ANOVA

Tests of Between-Subjects Effects

Dependent Variable: Traditional role portrayal

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	46.958[a]	5	9.392	10.712	.000	.691
Intercept	182.001	1	182.001	207.589	.000	.896
Programs	16.002	2	8.001	9.126	.001	.432
Gender	10.744	1	10.744	12.255	.002	.338
Programs * Gender	11.502	2	5.751	6.560	.005	.353
Error	21.042	24	.877			
Total	338.000	30				
Corrected Total	68.000	29				

a. R Squared = .691 (Adjusted R Squared = .626)

To help me understand how the two predictors interact, I will refer to the plot that displays the cell means, which is essentially the same information found in the cross-tabulation (see Table 10). Although the cross-tabulation table is informative on its own, the plot (Figure 8) helps us visualize the differences between the means by gender and program types and better understand the interaction between the two predictors.

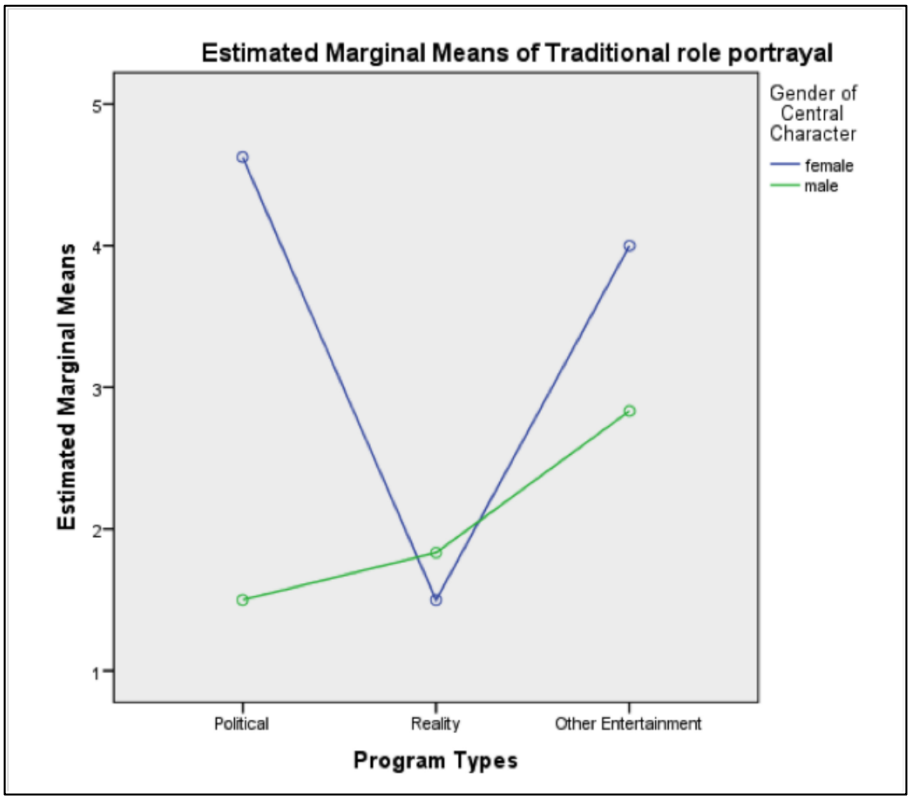


Figure 8. *The Plot of the Interaction Between Programs and Gender*

The blue line represents the means of traditional portrayal of female characters during ***political, reality,*** and ***other entertainment*** programs. Similarly, the green line represents the means traditional role portrayal of male characters in ***political, reality,*** and ***other entertainment*** programs. The y-axis represents the values of the traditional role portrayal, ranging from 1 (the lowest) to 5 (the highest), and the x-axis is our three program groups. The visual inspection tells me that during political programs, female characters are portrayed most traditional, while male characters are portrayed the least traditionally. A similar pattern seems to prevail in the TV commercials aired during other entertainment programs. Finally, there seems to be a reverse pattern when it comes to reality programs. The male characters tend to be portrayed more traditionally than the female characters. However, this difference does not appear to be as large as between the other two program types. In other words, the effect of program type doesn't appear to have the same effect on the male and female characters.

Notice that I interpret this plot with caution because further post hoc analyses can only confirm my predictions. In other words, just like the cross-tabulated table representing the means of our main factors, the plot only displays the patterns, which were not statistically tested, and further post hoc analysis is required to confirm my speculations. Further post hoc ***simple effects*** tests would compare the effect of gender at each level (type) of programs, as well as the effect of programs at each gender to see if there are any statistically significant differences. These comparisons require more advanced knowledge of statistics and more familiarity with the SPSS program, and so, we won't cover this topic in this beginner course. However, in your results, and in the discussion section, you can discuss the possibility of what further post hoc analyses could reveal, but also emphasize the need for further analyses to confirm your speculations.

Reporting Your Results:

When reporting results of the 2-way ANOVA, report the F-statistic (for the main effect of both independent variables and the interaction effect), the df (degrees of freedom for each factor and the error term, found in Table 11), and p-value. These statistics can be presented as follows: A two-way ANOVA was conducted to test the hypothesis that the traditional portrayal of main characters in TV commercials will depend on two factors, their gender, and the program type during which these commercials are aired. Although the main effects of both, gender and programs, were significant, $F(1, 24) = 12.25$, $p<.005$ and="" i="">$F(2, 24) = 9.13$, $p=.001$, respectively, the interaction effect of the two factors was also significant, $F(2, 24) = 6.56$, $p=.005$. The visual inspection of the plot of the estimated marginal means of traditional role portrayal by program and gender suggests that female characters are portrayed most traditionally in TV commercials during political programs and the least traditionally in reality programs. Conversely, male characters are portrayed least traditionally in political and most traditionally in other entertainment programs. However, simple effect comparisons are needed to confirm the statistical significance of these differences.

Paired Samples t Test

So far we have been comparing the means of the groups or levels that are considered to be independent of each other (e.g., political vs. entertainment programs). However, in other cases, instead of comparing the scores of two independent groups, a researcher may need to test difference in means of scores of either related or the same groups of people. For example, if a psychologist wants to test a new treatment, he or she may first measure the behavior before administering the treatment and, then, after treatment. The scores before and after the treatment are considered to be related, since they came from the same individuals. In other cases, the scores may not come from the same but somehow related people. These groups could be husbands and wives, two siblings, or two friends. In all of these cases, these are ***paired samples***.

Let's use a hypothetical study as an example to show how to perform this test. I will use fictitious data, Aggression paired samples. Suppose I conducted an experiment to find out if playing violent video games can cause aggressive behavior. I have a sample of 15 young people whose aggressive behavior was measured using some scale ranging from '1' (the lowest level of aggression) to '5' (the highest level of aggression). Each participant's behavior was measured twice, before they played video games (pretest) and after they played (posttest). As a result, I have two dependent variables, aggrs_pre (pretest score) and aggrs_post (posttest score). The question is, do these scores differ enough to allow me to conclude that playing video games causes aggression? To answer this question, I will conduct a paired samples t test. Here are the steps:

- I click on Analyze→Compare means→Paired Samples T Test
- I select my two variables, the means of which I want to compare, Aggrs_pre and aggrs_post. The score for pretest has to go first to variable 1, and the score for my posttest has to go to variable 2. If I had additional an pair of variables, I could have added them as my second pair.
- Click Ok

Now, let's interpret the results. The first table in the output (see Table 12) gives me the means of aggressive behaviors before and after playing the video games. These are the means that are being compared and tested.

Table 12. SPSS Output of the Means of Aggressive Behavior Before and After Playing Video Games

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	aggressive behavior before video games	2.60	15	1.121	.289
	aggressive behavior after video games	3.93	15	.961	.248

The next table (see Table 13) displays the correlations between the two paired scores. The correlation between the pre- and post-scores is significant, (r=.64), confirming the dependence of the two variables which is not surprising, since these are the scores of the same subjects but before and after playing the games.

*Table 13.*SPSS Output Table of the Correlation between Aggressive Behavior Scores (Before and After)

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	aggressive behavior before video games & aggressive behavior after video games	15	.636	.011

The third, key table (Table 14) in the output displays the results of the paired samples *t* test. The Mean column is the difference in points between before and after video games, which is 1.33 (it doesn't carry any meaning in this example). This difference is also statistically significant: t (14) = -5.74. Since the descriptive statistics table tells me that the post scores are higher than the prior scores, I can conclude that playing video games *did* increase the level of aggressive behavior.[11]

[11] Since this was invented data, no real conclusions can be made.

Table 14. SPSS Output of the Paired Samples *t* Test

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	aggressive behavior before video games - aggressive behavior after video games	-1.333	.900	.232	-1.832	-.835	-5.739	14	.000

Reporting Your Results:

Since my main goal was to compare the means of aggression before and after video game playing, I need to report the means and the standard deviations for both groups. The report should also indicate whether the difference was significant, by reporting the t-statistic, df, and the p-value. Thus, I would report these results as follows: *A paired-samples t-test was performed to test the group differences between before- and after-video game conditions. The test showed that the aggression level was higher after video game play (M=3.93, SD=.96) than before the video game play (M=2.60, SD=1.12), and the difference was statistically significantly, t(14)=-5.74, p< .001.*

9.4 Non-Parametric Tests for Questions of Group Differences and Association

Chi-square, Phi, and Cramer's V

The statistical analyses we have covered up to this point were essentially testing hypotheses about some parameters in a ***normally*** distributed population. Since my variables were drawn from a normally distributed population, the variables were also assumed to have a normal distribution; and my dependent variable was measured on either an interval or a ratio scale of measurement. But what if I cannot make the assumption of normality, i.e., that my data is normally distributed. Or suppose, my dependent variable is measured using an ordinal or a nominal scale of measurement: men and women, republicans and democrats, correct and incorrect responses, etc.? If one or both of these characteristics are present, you will have to conduct a *nonparametric test*, where no assumptions are made about the population of interest and your dependent variable can be nominal or ordinal. Keep in mind that because nonparametric tests are not based on the assumption of normality, i.e., variance is not measured to determine significance, they are less powerful in terms of what conclusions one can draw from the results of a nonparametric test.

There are several nonparametric tests available; a ***chi-square*** test is appropriate when we need to test difference between groups with a categorical dependent variable or when we need to test the association between two categorical variables. In other words, unlike parametric tests where we have different statistics to test, research questions and hypotheses of differences and associations, the same statistics can apply to either one of them. Unlike a*t*test, a chi-square test compares observed and expected frequency counts at each level of two variables to compute the statistic. To determine the expected value for each cell (each level of one variable by each level of another variable), complete the following: E_i=(Row total) (Column total)/ (Grand total). For example, to compute the expected value for the first cell in the first row of contingency Table 1, you would do the following: $E_{1.1}$ = (8)(10)/(20). So, the expected value ($E_{1.1}$) is, 4. Although this test does not make any assumptions about the variance of the data, it does require the data to have a large enough sample to yield reliable results. Thus, if the observe values in the cells are too low, the expected values will also be too low. The rule of thumb is the following: if the expected values are less than 5, a more appropriate test is the Fisher's exact test. Let's use the same data set that we used in example 1, the TV commercials data, to illustrate the computations and interpretations of a non-parametric test.

Let us ask the same research question as in example 1: *Do **political** and **entertainment** TV commercials differ in a stereotypically traditional portrayal of their main characters?* But this time my variable, traditional role portrayal is coded (i.e., measured) differently. This time the computations are conducted using a nominal scale of measurement: 'present' (i.e., stereotypically traditional ***family*** role portrayal) or 'absent' (i.e., not stereotypically traditional ***family*** role portrayal). Additionally, the stereotypical family role portrayal is operationally defined as the following: men are portrayed as breadwinners and women are portrayed as homemakers. Although my research question has not changed, I will have to use a chi-square test, because my dependent variable is now nominal. So, the null and the alternative hypotheses will be stated as follows:

H_0: Programs and traditional role portrayal of the main characters in TV commercials are independent of each other.

H_1: Programs and traditional role portrayal of the main characters in TV commercials are NOT independent of each other.

Notice that both hypotheses are phrased as *association statements/questions*. This is because we use chi-square to test both types of hypotheses.

Let's look at the data presented in a table to better understand the computations involved in this statistical test.

The first step is to compute the four expected values for each cell in my contingency table (see Table 15): 1.The family role present in TV commercials for political programs, 2. The family role present in TV commercials for entertainment programs, 3. The family role absent in TV

commercials for political programs and 4. The family role absent in TV commercials for entertainment programs. SPSS calculates this, and all subsequent computations, automatically for us.

Table 15. Contingency Table

Family Roles * Program Types Crosstabulation

Count

		Program Types		Total
		Political	Entertainment	
Family Roles	Absent	8	0	8
	Present	2	10	12
Total		10	10	20

Here are the steps:

- Choose Analyze → Descriptive Statistics → Crosstabs
- Select Family Roles and move it to the Rows box. Move Program Types to the Columns box (see Figure 9).

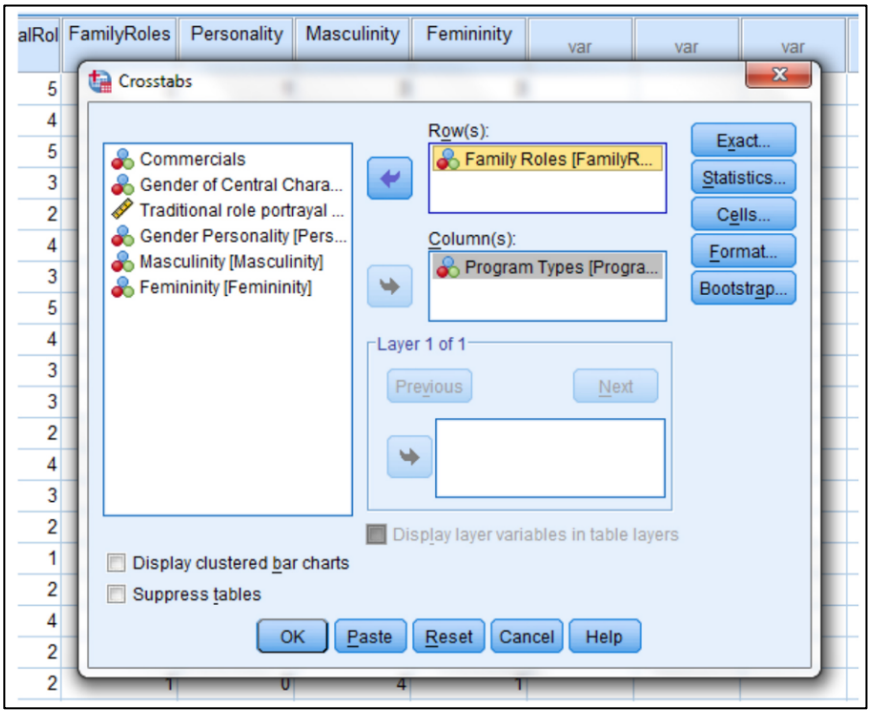


Figure 9. *Chi-Squire Test: Crosstabs.*

- Click on Statistics → Select Chi-square and Phi and Cramer's V[12]
- Click on Continue → choose Cells → select Expected, Observed, and Total (see Figure 10) → Continue and Ok.

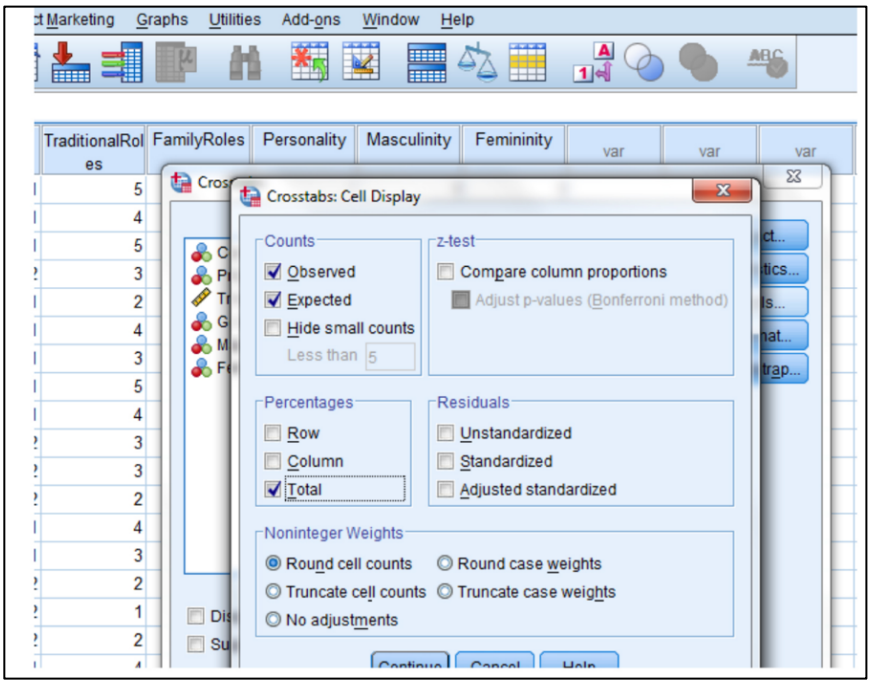


Figure 10. *Chi-Squire Test: Cells.*

Let's first look at the table that gives us the observed and expected values (see Table 16). Notice that the expected value of the first cell in the first row is four, which is exactly what we got when we calculated it by hand.

Table 16. Observed and Expected Values of Family Roles and Gender Crosstabulated

Family Roles * Program Types Crosstabulation

			Program Types		
			Political	Entertainment	Total
Family Roles	Absent	Count	8	0	8
		Expected Count	4.0	4.0	8.0
		% of Total	40.0%	0.0%	40.0%
	Present	Count	2	10	12
		Expected Count	6.0	6.0	12.0
		% of Total	10.0%	50.0%	60.0%
Total		Count	10	10	20
		Expected Count	10.0	10.0	20.0
		% of Total	50.0%	50.0%	100.0%

[12] Phi and Cramer's V statistics will be discussed next.

This table tells us that the expected count, of the TV commercials that did not portray their main characters in stereotypical, family roles during political programs, is 4 and the observed or actual count is 8. Thus, there are 4 more TV commercials than would be expected by chance, given the totals shown in the table. Furthermore, we can see that there are 4 less than expected TV commercials (i.e., zero) that did not portray their main characters in the stereotypically familial roles during the entertainment programs. Some differences between the expected and the observed values, are also found for the TV commercials aired during political and entertainment programs, which portray their main characters in stereotypical family roles. What these descriptive results suggest is that, unlike what I would expect, more TV commercials aired during entertainment programs have their main characters portrayed in stereotypical, familial roles. However, what is still unknown is if these differences are statistically significant, or are not simply due to pure chance. To determine if the difference between the two groups is indeed statistically significant, i.e., big enough to not occur by chance, we need to turn to the chi-squire statistic, displayed in Table 17.

But first, notice that the expected values in our cells are too low, less than 5. Luckily, SPSS performs both, the chi-square and the Fisher's exact test, for a 2x2 table. Since my data is too small, I will refer to the results of the Fisher's exact test (see Table 17).

Table 17. Results of the Chi-square and Fisher's Exact test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	13.333[a]	1	.000		
Continuity Correction[b]	10.208	1	.001		
Likelihood Ratio	16.912	1	.000		
Fisher's Exact Test				.001	.000
Linear-by-Linear Association	12.667	1	.000		
N of Valid Cases	20				

a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 4.00.

b. Computed only for a 2x2 table

This table tells us that the expected count, of the TV commercials that did not portray their main characters in stereotypical, family roles during political programs, is 4 and the observed or actual count is 8. Thus, there are 4 more TV commercials than would be expected by chance, given the totals shown in the table. Furthermore, we can see that there are 4 less than expected TV commercials (i.e., zero) that did not portray their main characters in the stereotypically familial roles during the entertainment programs. Some differences between the expected and the observed values, are also found for the TV commercials aired during political and entertainment programs, which portray their main characters in stereotypical family roles. What these descriptive results

suggest is that, unlike what I would expect, more TV commercials aired during entertainment programs have their main characters portrayed in stereotypical, familial roles. However, what is still unknown is if these differences are statistically significant, or are not simply due to pure chance. To determine if the difference between the two groups is indeed statistically significant, i.e., big enough to not occur by chance, we need to turn to the chi-squire statistic, displayed in Table 17.

But first, notice that the expected values in our cells are too low, less than 5. Luckily, SPSS performs both, the chi-square and the Fisher's exact test, for a 2x2 table. Since my data is too small, I will refer to the results of the Fisher's exact test (see Table 17).

Table 17. Results of the Chi-square and Fisher's Exact test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	13.333[a]	1	.000		
Continuity Correction[b]	10.208	1	.001		
Likelihood Ratio	16.912	1	.000		
Fisher's Exact Test				.001	.000
Linear-by-Linear Association	12.667	1	.000		
N of Valid Cases	20				

a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 4.00.

b. Computed only for a 2x2 table

According to Fisher's exact test, the differences between the expected and the observed values are significant, and so we can reject the null hypothesis and conclude that the TV commercials aired during political programs in my sample are different from the ones aired during the entertainment programs in terms of how stereotypically they portray their main characters. Table 16 helps me in further interpretations, which is, opposite of what I expected. More TV commercials that were aired during entertainment programs portrayed their main characters in stereotypically traditional roles as compared to the commercials of political programs.

Reporting Your Results:

The following statistical information should be reported when we communicate these results in a research paper: A chi-square test showed that more TV commercials which were aired during entertainment, as opposed to, political programs, portrayed their main characters in stereotypically

familial roles. This difference was statistically significant. (χ^2 =13.33[13], *df* = 1, *p*<.001, if I were to report the chi-square test), or *p* <.001 i=""> if we were to report Fisher's exact test; although both tests, in this case, yield similar results).

Limitation of chi-square test and the solution with Phi and Cramer's V:

One important limitation of a chi-square test is that it can only tell us whether the difference between the expected and the observed values is significant, or whether or not the variables are significantly related. However, the magnitude or the strength of this relationship remains unknown. In other words, the absolute value of a chi-square statistic of 13.33 doesn't have any meaning. If I were to know how strongly my variables are associated, I will have to conduct a Phi (for a 2 x 2 cross-tabulation like the one in our example) or Cramer's V (for a 3 x 3 cross-tabulation) test. SPSS can automatically calculate both statistics if you select them along with a chi-square test. Table 18 displays the results of Phi and Cramer's V tests.

*Table 18.*Results of Phi and Cramer's V Test

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.816	.000
	Cramer's V	.816	.000
N of Valid Cases		20	

Because my data is a 2 by 2 cross-tabulation, I have to refer to the phi statistic and report it as phi = .816, *p* = .000. Since we already know that Family Roles and Programs Types are significantly associated, what we are interested to find out from the phi coefficient is the magnitude (strength) of this association. Its interpretation is almost identical to how we would interpret a correlation coefficient (see the chapter on descriptive statistics and correlations on how to interpret a correlation coefficient).

[13] Since my values are too small, I will have to report Fisher's exact test.

9.5 Parametric Test of Questions of Strength of Association (Relationship)

Research questions that fall within the category of association are questions about whether and to what extent two or more variables are related. The following are just a few examples of how these questions are usually stated:

- Does playing violent video games contribute to aggressive behavior? (question of association between playing violent video games and aggression)
- Has there been a reduction in the stereotypical portrayal of women in TV commercials over time? (question of association between women's stereotypical images and time)
- Does red wine have any health benefits? (question of association between red wine and health)
- Can physical exercise reduce depression? (question of association between exercise and depression)

To choose an appropriate inferential statistic to test associations between variables: The following two factors will have to be taken into consideration when choosing statistical analysis to test your hypothesis: (1) the scale of your variables , and (2) the number of independent/predictor variables (see Table 19).

Table 19. Inferential Statistic for Research Questions of Association

Two variables in the analysis: 1 independent/predictor variable and 1 dependent/criterion variable			**3 or more variables: 2 or more independent or predictor variables and 1 dependent/criterion variable**
NON-PARAMETRIC STATISTICS		**PARAMETRIC STATISTICS**	
One or both variables is/are *nominal*	**Both or one variable is/are *ordinal***	**Both variables are *scale***	**Independent variables are scale and/or dichotomous, dependent/criterion variable is *scale***
Phi or Cramer's V	**Spearman (rho)**	**Pearson (r)**	**Multiple Regression**

**This table was adapted from Morgan, Griego, & Gloeckner (2001).*

We have already gone over the computations of Phi and Cramer's V statistics in the previous section. Recall, that when the dependent variable is nominal or ordinal, and when we are interested to know the strength of the relationship between the variables, regardless whether it is a question of group differences or association, Phi or Cramer's V statistics are appropriate to address both research questions. The computations of the Pearson (r) and Spearman (rho) statistics have also been covered in the chapter on descriptive statistics. So in this section, I will demonstrate how to compute and interpret simple and multiple linear regression tests.

Multiple (and Simple) Linear Regression

Suppose I have a simple hypothesis that police funding predicts the crime rate in a city. To test this hypothesis of ***association***, I will have to do a *linear regression test*. The idea behind a linear regression test is that we use one or more variables to predict another variable. This prediction is based on finding the regression line that best describes the relationship between those two (or more) variables. To represent a possible relationship between police funding and the crime rate,[14] we can draw a scatter plot where police funding data (measured in dollars per resident) is plotted on the x-axis, and the crime rate (measured as the total overall reported crime rate per 1 million residents) is plotted on the y-axis (see Figure 11). The line that you see in the scatter plot is the best-fitted (regression) line that models the relationship between these two variables; it is found based on the method of least-squares.

We won't cover the calculation or the principle behind this method, but I will show you, later on in this chapter, how we can use SPSS to calculate it. Also notice that the regression model is the equation that is being represented in the scatter plot: Y=2.9E2 + 11.34*X. We will come back to the regression equation later in this section.

[14] This is real data (Crime.sav), taken from college.cengage.com. See Reference page for the full website link and access the data.

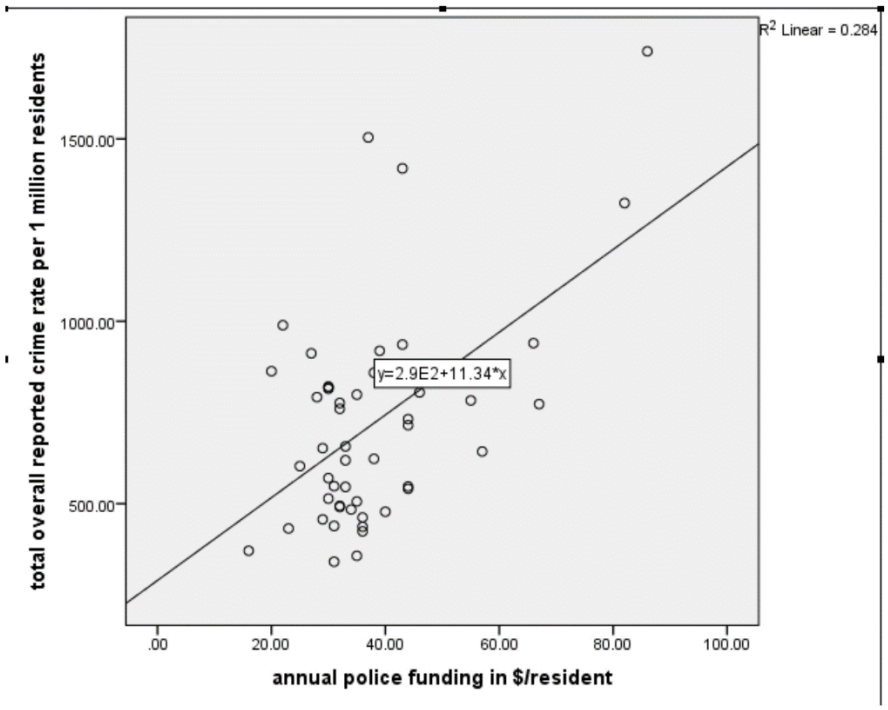


Figure 11. *The Scatter plot with the regression line representing an association between police funding and reported crime.*

Important Assumptions

Besides the requirement to have a dependent variable measured on an interval or a ratio scale, there are several other important assumptions that have to be met: a). the relationships between variables have to be linear, b). all observations have to be independent, c). the data has to be homoscedastic (similar variance along the regression line), and d). no significant outliers. Ideally, every researcher needs to evaluate his or her data first to make sure that it meets all of the mentioned assumptions before conducting multiple regression. But this topic goes beyond the scope of this chapter and will not be covered. For now, let's assume that our data have met all the assumptions and we can proceed with the testing.

The Xs in regression are the predictor variables that are supposed to predict or to explain the Y criterion, or dependent variable. To let SPSS compute a simple linear regression (with only one predictor variable), I will go through the following steps:

Analyze → Regression → Linear → move my dependent variable, Total overall reported crime rate to **Dependent** box and my predictor, Annual police funding to **Independent(s) box**

Click OK.

The SPSS will produce these two important tables (see below). The ANOVA) table shows whether the overall model with all of the predictors is significant. In this case, there is only one predictor, and it is, in fact significant as the p-value is lower than 0.05 (highlighted in yellow).

The SPSS will produce these two important tables (see below). The ANOVA) table shows whether the overall model with all of the predictors is significant. In this case, there is only one predictor, and it is, in fact significant as the p-value is lower than 0.05 (highlighted in yellow).

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1203611.943	1	1203611.943	19.067	.000[b]
	Residual	3029987.977	48	63124.750		
	Total	4233599.920	49			

a. Dependent Variable: **total overall reported crime rate per 1 million residents**
b. Predictors: (Constant), annual police funding in $/resident

The second table (see below) lists the statistics of each individual predictor in the model. In this case, it is only one predictor, and, it is a significant predictor.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	289.749	104.304		2.778	.008
	annual police funding in $/resident	11.340	2.597	.533	4.367	.000

a. Dependent Variable: total overall reported crime rate per 1 million residents

Suppose I want to test a model that has two predictors of the crime rate, police funding and the percentage of young people (16 to 19 year olds) who are not in high school or high school

graduates. My specific question is to know which one of the two factors will be a significant predictor of crime in a city? Now, I will compute a ***multiple*** linear regression with more than one predictor. However, the computational steps are the same as for a simple regression (Analyze --->Regression --> Linear). In addition to police funding, I will also enter my second variable, the percentage of young people not in high school or college. Here are the results:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1318320.509	2	659160.254	10.627	.000[b]
	Residual	2915279.411	47	62027.222		
	Total	4233599.920	49			

a. Dependent Variable: total overall reported crime rate per 1 million residents
b. Predictors: (Constant), % of 16 to 19 year-olds not in highschool and not highschool graduates., annual police funding in $/resident

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	202.915	121.521		1.670	.102
	annual police funding in $/resident	10.193	2.709	.479	3.762	.000
	% of 16 to 19 year-olds not in high school and not high school graduates.	8.453	6.216	.173	1.360	.180

a. Dependent Variable: **total overall reported crime rate per 1 million residents**

The conclusion is that police funding, again, is the only significant predictor of the crime (see the p-values for both predictors).

Another Example of a Multiple Linear Regression Analysis when Predictor Variables are Categorical

To help me cover this topic I will use the TV commercials data set (TVCommercials.sav) with which you are already familiar. Suppose I hypothesize that the traditional portrayal of TV commercials' main characters is *related* to the programs during which they are aired, e.g., entertainment, political, etc., and also to the gender of the main characters. In other words, I believe that not only political programming will produce TV commercials with more traditionally portrayed characters, but also that the female characters in those same commercials will be more stereotypically portrayed.

My criterion variable Y is Traditional Roles (i.e., the interval scale, the degree or the extent of traditional portrayal). If the test shows that my X_1 and X_2 predict Y, this will suggest that a significant relationship exists between those variables, thus, supporting my hypothesis. This is an example of a ***multiple*** linear regression (multiple because the model has more than one predictor variables). A hypothesis with only one predictor variable would require a simple linear regression test (simple, because the model would include only one predictor). Conducting and interpreting simple linear regression is easier once you know how to perform and interpret multiple linear regression.

Dummy Coding

Before computing linear regression, I will need to make a small coding transformation of my predictor variables. The nominal codes need to be changed from '1' (political programs) and '2' (entertainment programs) to '1' and '0', respectively, and from '1' (female) and '2' (male) to '1' and '0', respectively. This new coding is called *dummy coding*, i.e., using only ones and zeros. This coding makes the interpretation of the results much easier at the end. Dummy coding can be also applied to variables with more than 2 levels. For example, suppose I had three levels of programs: '1' *political*, '2' *entertainment*, and '3' *other*. Here is how this hypothetical data set would look like in SPSS (see Figure 12):

	Commercials	Programs	var
1	1	1	
2	2	2	
3	3	1	
4	4	1	
5	5	3	
6	6	1	
7	7	1	
8	8	1	
9	9	3	
10	10	3	
11	11	2	
12	12	2	
13	13	3	
14	14	2	
15	15	3	
16	16	3	
17	17	2	
18	18	3	
19	19	2	
20	20	3	
21	21	1	
22	22	1	
23	23	3	

Figure 12. *Data set with the nominal three-level predictor variable Programs.*

To transform a nominal variable with k-levels into a dummy variable, I will create a set of k-1 dichotomous variables containing the same information. For example, my hypothetical 3-level variable will become a set of 2 (i.e., 3-1) dichotomous variables (see Table 20):

Table 20

Programs	Political	Entertainment
Political (1)	1	0
Entertainment (2)	0	1
Other (3)	0	0

This transformation can be easily completed in SPSS, just follow these steps. For an illustration let's first create a new dummy variable, *political*, from my original variable, *programs* (nominal variable):

- Transform→Recode into Different Variables
- Move Programs to Numeric Variable/Output Variable box and name this newly constructed dummy variable ***Political*** → Change (see Figure 13).

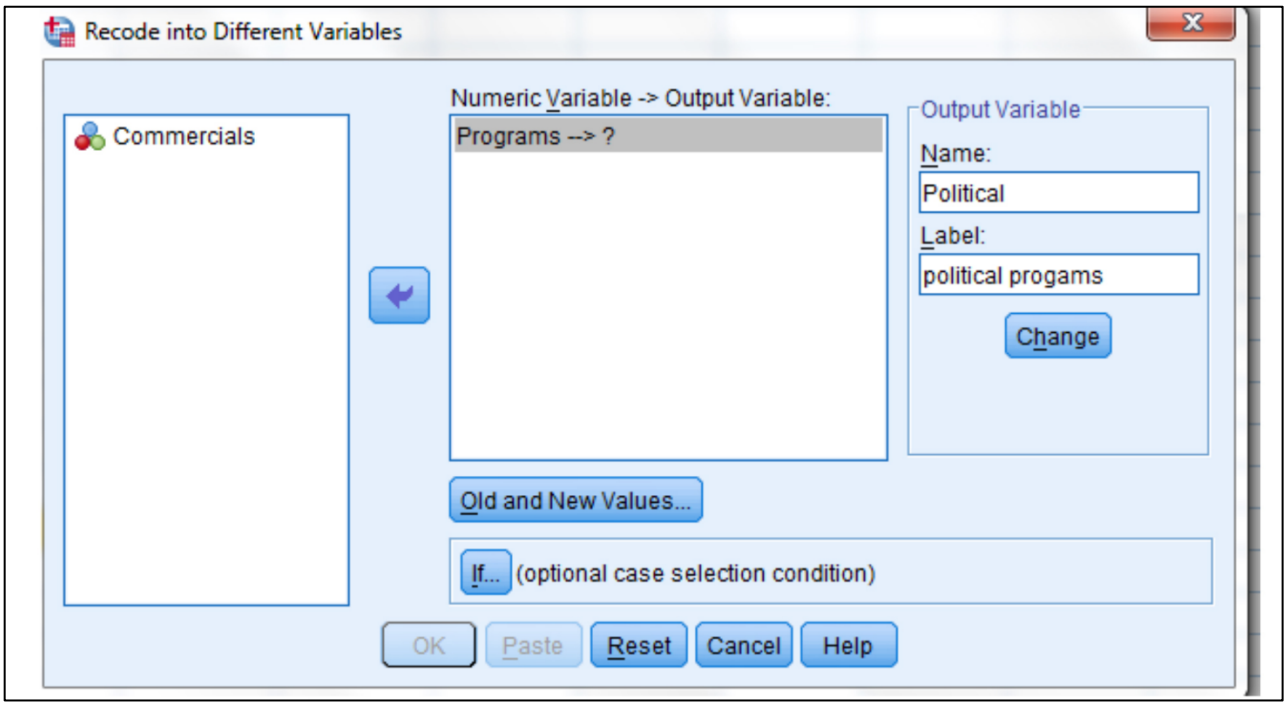


Figure 13. *SPSS numeric variable/output variable box.*

Next, I need to give new codes to my program levels. I follow the coding scheme outlined in Table 20, where political programs are coded as '1', and the other two types of programs are given '0s'.

- Click on, Old and New Values (to indicate what you want your new values to be)
- I typed 1 in the Old Value window and 1 in the New Value window since my political programs remains coded as '1' → click Add
- I typed, 2 in the Old Values window and, 0 in the New Value window to change my entertainment programs' code to '0' → Add
- I typed, 3 in the Old Values window and, 0 in the New Value window to change my other programs also to '0' → Add
- Continue (see Figure 14)

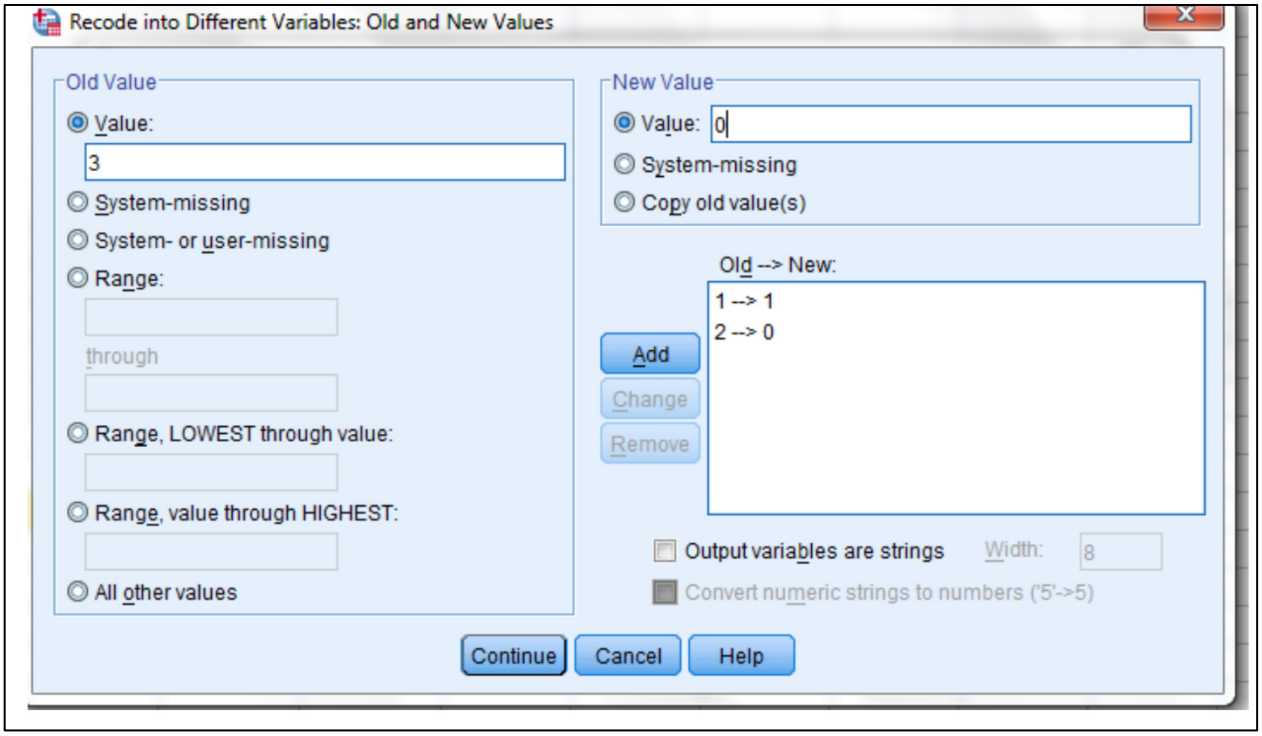


Figure 14. *SPSS recoding Political into dummy codes.*

Follow similar steps to create the second dummy variable, *Entertainment*. This time, entertainment programs will be coded as '1', and the other two are coded as '0' (see Figure 15, 16).

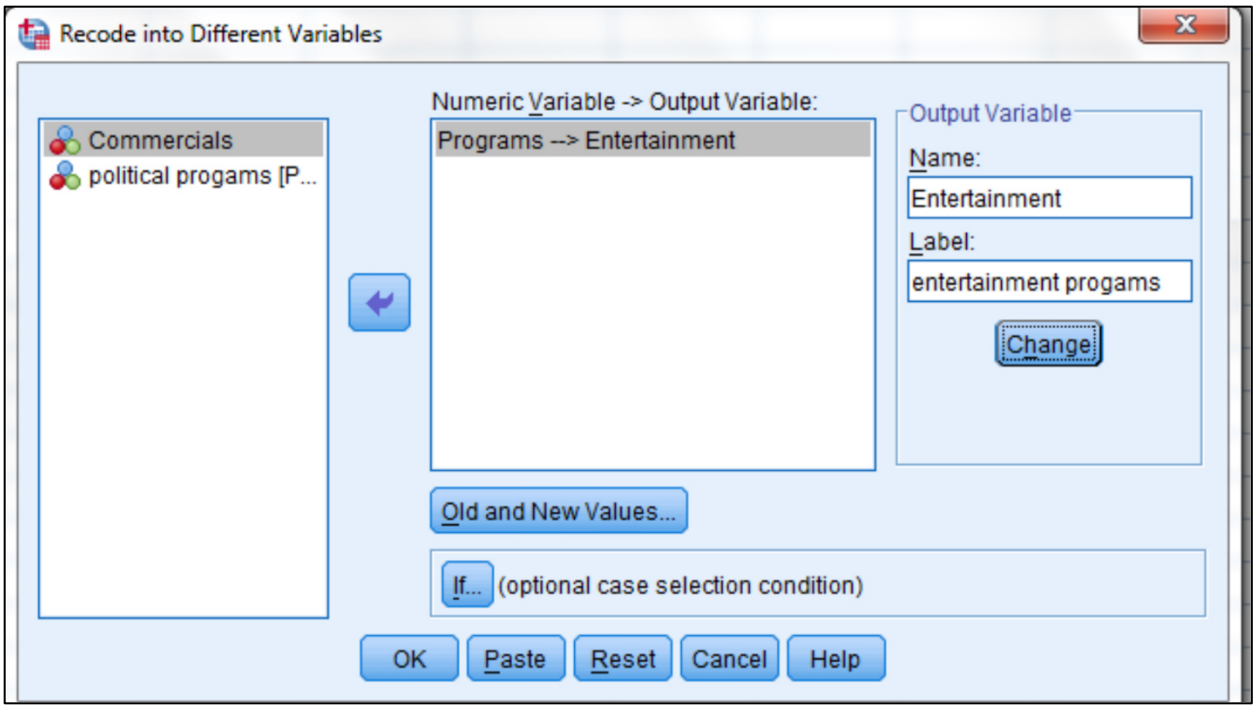


Figure 15. *Creating the dummy variable, Entertainment.*

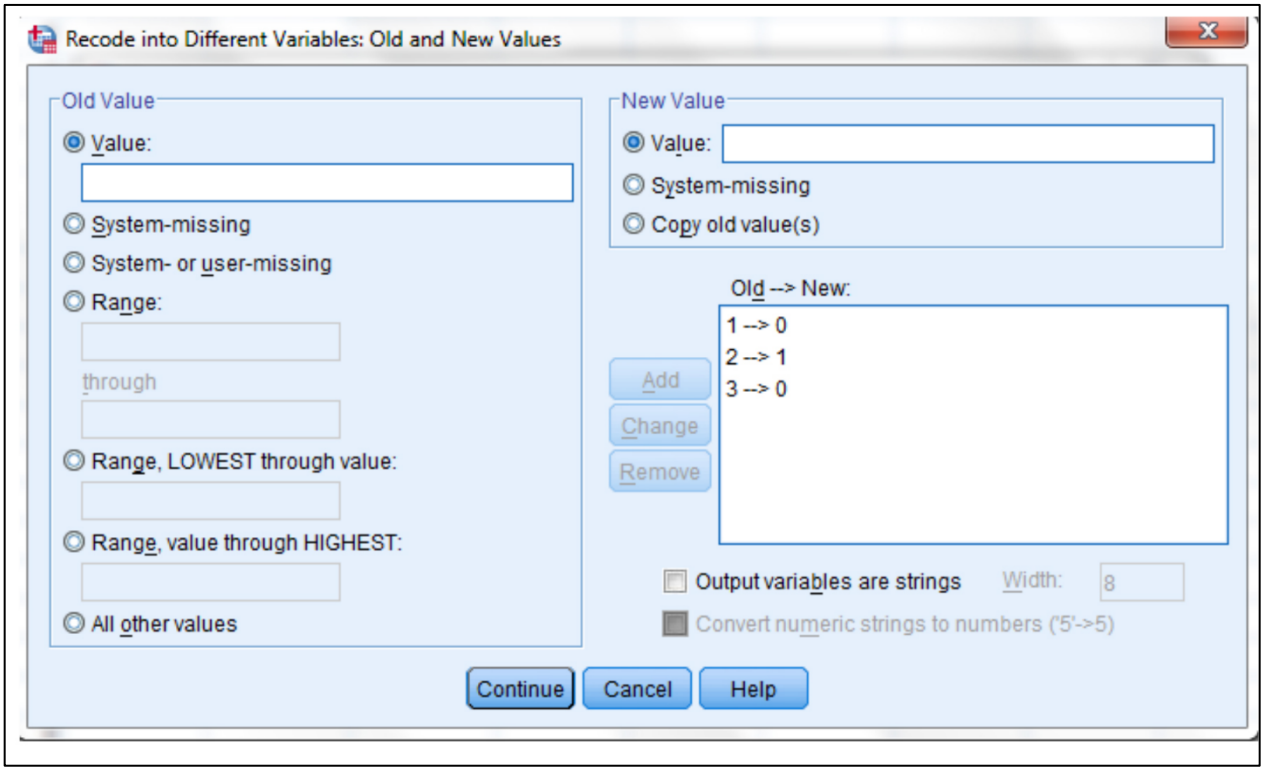


Figure 16. *Dummy codes for Entertainment Programs.*

	Commercials	Programs	Political	Entertainment
1	1	1	1	0
2	2	2	0	1
3	3	1	1	0
4	4	1	1	0
5	5	3	0	0
6	6	1	1	0
7	7	1	1	0
8	8	1	1	0
9	9	3	0	0
10	10	3	0	0
11	11	2	0	1
12	12	2	0	1
13	13	3	0	0
14	14	2	0	1
15	15	3	0	0
16	16	3	0	0
17	17	2	0	1
18	18	3	0	0

Figure 17. *Dummy coded, three-level, variable Programs.*

To conduct a multiple regression test, both dummy variables, *Political* and *Entertainment*, have to be included in the model. The reason why we don't need to create a third variable for the *Other* category is, because its information is already included in the model, and is marked as '0' in both Political and Entertainment variables (see Table 20 and Figure 17).

Let's go back to my original, two-level predictor variables, *Programs* and *Gender*. My new dummy coded predictors are, *ProgramsR*[15] and *GenderR*. Figures 18 and 19 show the steps to renaming and recoding, *Programs*.

[15] Since my *Predictor Variable Program* has only two levels, I only need to create one variable (k-1),which I named, ProgramsR. Similarly, my variable, Gender, has only two levels and I created only one dummy variable—GenderR.

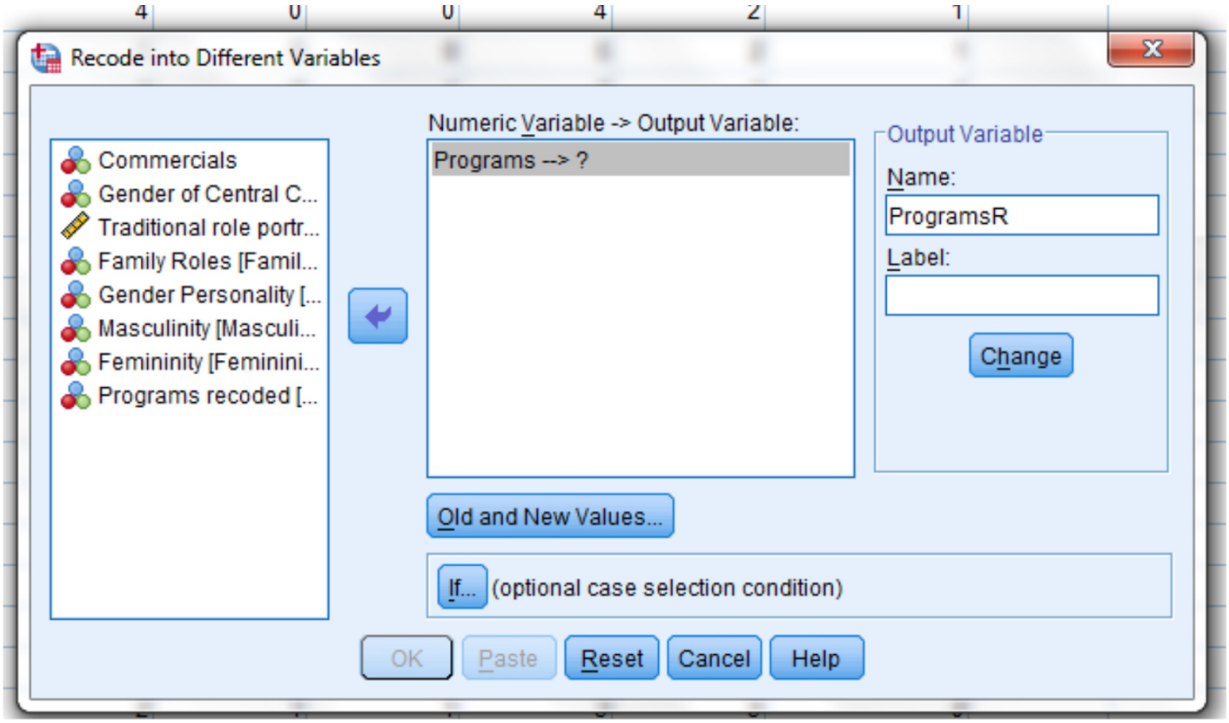


Figure 18. *Naming newly created dummy variable Program as ProgramsR.*

The code '1' for political programs will remain as '1', but my code '2' for entertainment will be changed to a '0' (see Figure 19).

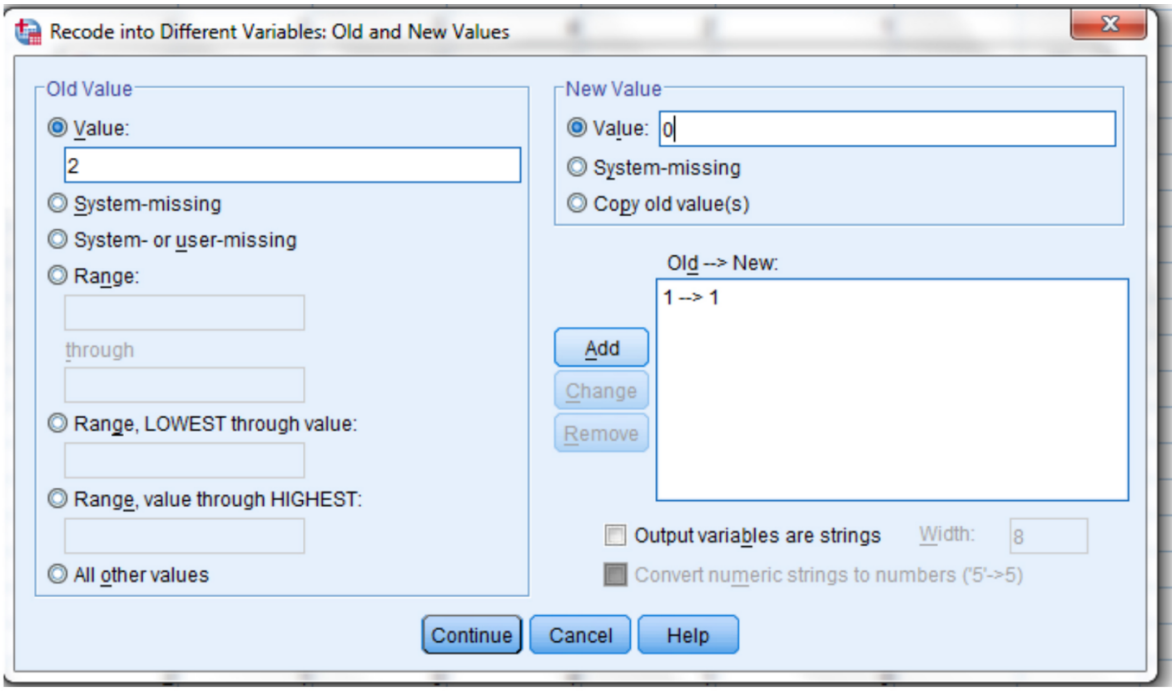


Figure 19. *Assigning dummy codes to ProgramsR.*

Similarly, female code, '1' remains as '1', and the male code, '2' has been changed to a, '0' in *GenderR*. I could have just as easily assigned male to be, '1' and female to be, '0'. Although, the choice of assigning codes is somewhat arbitrary, it will be an important piece of information to keep in mind when we have to interpret our results. We will come back to this point in the interpretation part of this section.

Now, we are ready to perform the multiple linear regression test and to test our hypothesis. Here are the steps:

- Click on Analyze→Regression→Linear
- Move criterion variable *Traditional role portrayal* to the Dependent variable box and the two predictor variables, *Gender* and *Programs*, to the Independent variable box (see Figure x).
- *Method* should remain as default, *Enter*.
- Click on Statistics→Select, Descriptives and ensure that Estimates and Model fit are also selected (see Figures 20-21).
- Continue and OK.

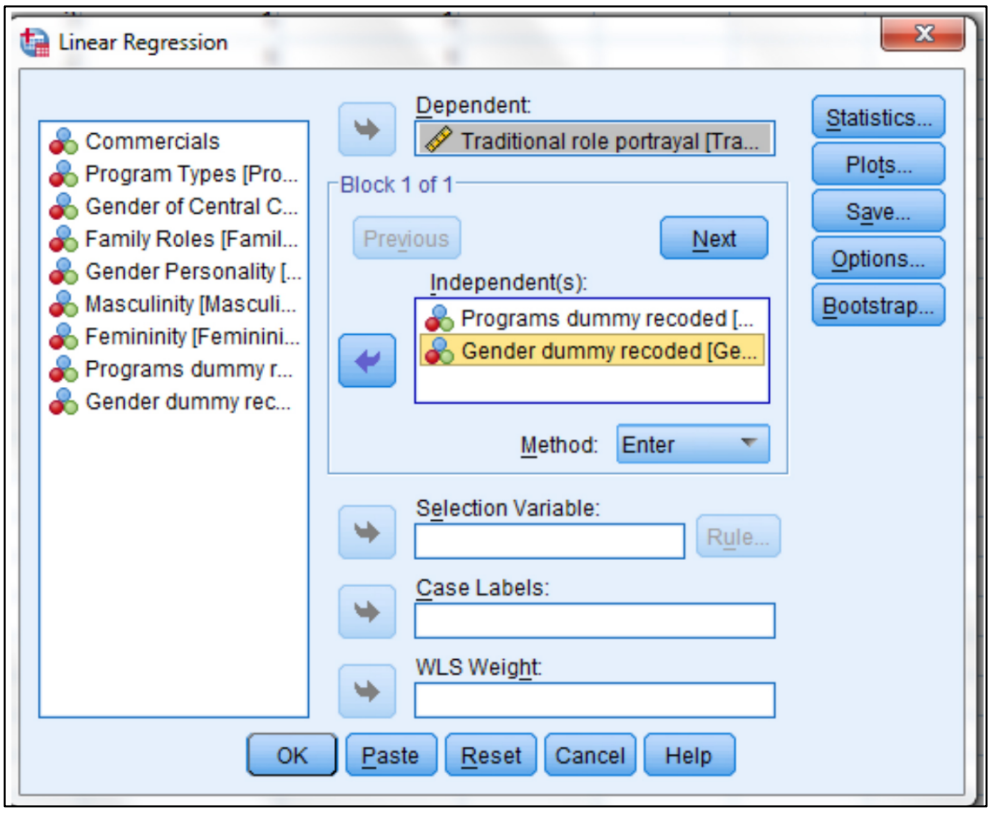


Figure 20. *Linear Regression: Variables.*

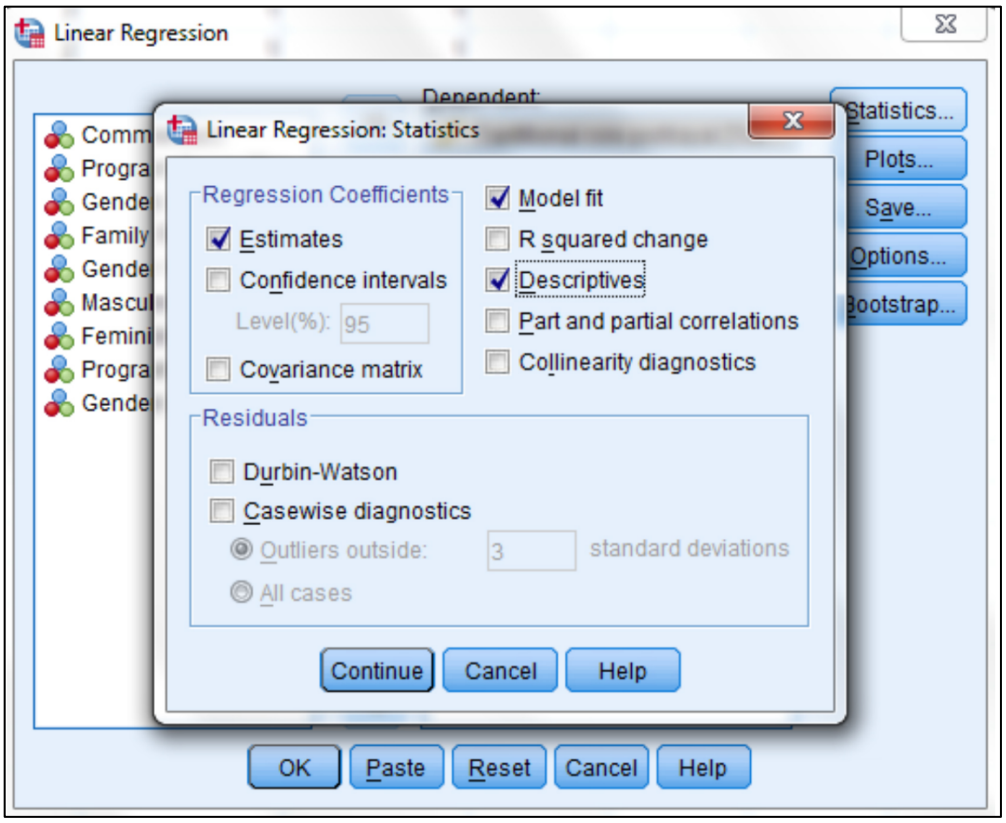


Figure 21. *Linear Regression: Statistics.*

Interpreting Results. The first two tables in the SPSS output are the descriptive statistics, the means, the standard deviations, and the correlations between the variables (see Tables 21, 22).

Table 21. SPSS Output of the Descriptive Statistics.

Descriptive Statistics

	Mean	Std. Deviation	N
Traditional role portrayal	3.15	1.182	20
Programs dummy recoded	.50	.513	20
Gender dummy recoded	.60	.503	20

Table 22 displays the bivariate correlations with traditional role portrayal. Both predictor variables are significantly correlated with this criterion variable (I circled their significance levels in red).

Table 22. Output of the SPSS Multiple Regression Test: Bivariate Correlations.

Correlations

		Traditional role portrayal	Programs dummy recoded	Gender dummy recoded
Pearson Correlation	Traditional role portrayal	1.000	.564	.638
	Programs dummy recoded	.564	1.000	.408
	Gender dummy recoded	.638	.408	1.000
Sig. (1-tailed)	Traditional role portrayal	.	.005	.001
	Programs dummy recoded	.005	.	.037
	Gender dummy recoded	.001	.037	.
N	Traditional role portrayal	20	20	20
	Programs dummy recoded	20	20	20
	Gender dummy recoded	20	20	20

Table 23 shows that the predictor variables were all simultaneously entered into the model.[16]

Table 23. Output of the SPSS Multiple Regression Test: Variables Entered/Removed

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Gender dummy recoded, Programs dummy recoded[b]	.	Enter

a. Dependent Variable: Traditional role portrayal

b. All requested variables entered.

[16] There are other methods of entering predictors into the model, but we will not be covering them in this book.

The model summary table (Table 24) provides you with the multiple correlation coefficient (R), which is the correlation coefficient when both predictors are entered simultaneously. It ranges between 0 and 1 (cannot be negative) and its value is interpreted similarly to a correlation coefficient value—the closer the value to 1, the higher the correlation or the magnitude of the association.

Table 24. SPSS Output of the Multiple Regression Test: Model Summary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.719[a]	.518	.461	.868

a. Predictors: (Constant), Gender dummy recoded, Programs dummy recoded

A more informative statistic is the adjusted ***R^2. This illustrates the percentage of variance in the dependent variable that is explained by the predictor variables.*** It is adjusted (increases or decreases) based on the number of predictor variables in the model that actually, significantly predict the dependent variable (we will come back to this point). Thus, the adjusted R^2 in our model tells us that 46% of the variance in the traditional role portrayal of the main characters is explained by one or both of my predictor variables.[17] The adjusted R^2 is also a rough estimate of the effect size, and 46% of the explained variance is a relatively large effect size.

Let's move on to the ANOVA table (Table 25), which gives us the *sums of squares*, *degrees of freedom*, *mean of squares*, and *F* statistic. The one you should be particularly interested to know is F statistic. This tells us whether our overall model is significant, that is, that at least one of my predictor variables significantly explains my criterion variable. To help you understand the meaning of this statistic, let's restate my hypothesis in terms of the null and alternative hypotheses:

H_0: The regression equation (with two predictors in the model) does not significantly predict the variance in my criterion variable (i.e., traditional role portrayal of the TV commercials' main characters, Y scores).

[17] To know what predictor variable is actually and significantly explaining my criterion variable, I have to look at the Coefficients table.

H_1: The regression equation (with two predictors in the model) significantly predicts the variance in the criterion variable (i.e., traditional role portrayal of the TV commercials' main characters, Y scores).

So, the *F* statistic tells us whether our alternative hypothesis can significantly explain the variance in my criterion variable. Given that *F* =9.12, and it is statistically significant (see Table 25), I can reject the null hypothesis and conclude that my regression model significantly predicts traditional role portrayal. Notice, however, that this doesn't mean that both of my predictors are equally contributing to the prediction; the overall model can be significant with only one significant predictor in the model. To know the individual contribution of every predictor in the model, we have to evaluate the significance of the slope, *b*, or the beta coefficient for each predictor in the model. This is covered in the next section.

Table 25. SPSS Output of the Multiple Regression Test: ANOVA

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13.740	2	6.870	9.117	.002[b]
	Residual	12.810	17	.754		
	Total	26.550	19			

a. Dependent Variable: Traditional role portrayal

b. Predictors: (Constant), Gender dummy recoded, Programs dummy recoded

Beta Coefficients

The coefficients table is the most informative table in the SPSS output (see Table 26). It lists all my predictors, their standardized and unstandardized beta coefficients, and whether or not they are significant (*t* and sig.). Let me elaborate more on the meaning of these statistics.

Recall that a linear regression test is based on finding the regression line that best describes the relationship between the variables of the study, i.e., independent/predictor variable and dependent/criterion variable. The equation of my linear regression model can be written as the following equation: $\mathbf{Y=a + b_1X_1 + b_2X_2}$, where **Y** is the value of my criterion variable (traditional role portrayal), a or alpha is the constant or the intercept, **b_1 is the slope or beta coefficient for X1** (program types), and **b2 is the slope or beta coefficient for X2** (gender of main characters).

Notice that the table gives unstandardized (usually denoted by the letter B) and standardized beta (usually denoted by the Greek letter β) coefficients. The difference between the two is that the

former is the relationship between the predictor variable and the criterion/dependent variable expressed in their original raw units (a one-unit change in the predictor/independent variable is associated with B units of change in the criterion/dependent variable), while the latter relationship is expressed when all the variables are converted into z-scores (a one-standard- deviation change in the predictor/independent variable is associated with β standard-deviation change in the criterion/dependent variable).

If you need to compare coefficients, measured in different units, and with different variance and mean, you have to use standardized coefficients. However, interpretation of the relationship between the predictor and the criterion variable is usually easier to interpret using unstandardized coefficients when the units, in which the variables are measured, are familiar to people, such as, years of education, degrees, percentages, etc.. The direction of the relationship is represented by the sign of the beta coefficient. So, for example, a positive beta coefficient represents the following relationship: a one-unit ***increase*** in the predictor/independent variable is associated with B units ***increase*** in the criterion/dependent variable. Similarly, the negative beta coefficient represents a B units ***decrease*** in the criterion/dependent variable with a one-unit increase in the predictor/independent variable.

Interpretation of the beta coefficients of dummy variables is slightly different. A one-unit difference represents a switch from one category/level to another, and B is the average difference in Y (criterion/dependent variable) between the category for which the predictor variable is 0 and the category that is 1. To properly interpret the slope of my predictor variable, *Programs*, we need to consider the following pieces of information: that my dummy code '1' stands for political programs, that its slope is .840 and is positive (see Table 26), and that my criterion variable (traditional role portrayal) ranges from 1 to 5 (in 1-unit increments). When considering all this information, we can conclude the following: that the traditional role portrayal of the main characters, on average, is .840 units more in political than entertainment programs, ***while holding my other predictor variable constant***. When interpreting the slope for gender, the traditional role portrayal of the main characters, on average is 1.150 units more when the main characters are female (females are coded as 1), and ***holding my other predictor variable constant***. If the sign were negative, the interpretation would be the following: traditional role portrayal of the main characters, on average, is 1.150 units less when the main characters are females (females are coded as 1 and holding the other predictor variable constant).

Thus, each beta coefficient represents an independent contribution of a given predictor/independent variable to the prediction of a given criterion/dependent variable, after the contribution of all other predictors in the model have been partialled out. This is an important difference between the bivariate correlations and the multiple regression results. For example, looking back at the *Bivariate Correlations* table (Table 22), both predictors are shown to be significantly correlated with our criterion variable. However, this is not the case when both predictors' contributions are considered in the regression model (see Table 26).

The beta coefficient for *Programs* (.840) is no longer significant, t=1.98, p = .065, but the beta coefficient for *Gender* remains to be significant (1.15), t = 2.65, p=0.017. This finding tells us that, while the types of programs the TV commercials are broadcast for is associated with the traditional portrayal of the characters in TV commercials; the genders of the main characters seem to be a better predictor. In other words, although the main characters in commercials seem to be more stereotypically portrayed in commercials during political programs, female characters are more likely to be stereotyped in TV commercials during political programs, and possibly during other programs. This important finding would not be known if we (1) did not include both variables in the model, and (2) if we only looked at the overall significance of the model (*F* statistic in Table 25).

Table 26. SPSS Output of the Multiple Regression Test: Coefficients

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.040	.325		6.281	.000
	Programs dummy recoded	.840	.425	.365	1.975	.065
	Gender dummy recoded	1.150	.434	.489	2.650	.017

a. Dependent Variable: Traditional role portrayal

Reporting Your Results:

Typically, you will need to report the R^2 and the adjusted R^2, F, and the independent contribution of each predictor variable in the model. Thus, I would report my results in the following manner: A multiple regression analysis was conducted with two predictor variables, programs, and gender. While the overall model was significant, R^2=.518, adjusted R^2= .461, F (2, 17) =9.12, p=.002, only gender, and not programs, made a significant contribution to the prediction of traditional role portrayal, β =1.15, p=.015 andβ =.840, p=.065, respectively.

Guided Practice (not graded)

Below is an SPSS output of a multiple regression analysis using the following variables:[18]

Name	Type	Width	Decimals	Label
ID	Numeric	8	0	movie
firstyrbxrec	Numeric	8	0	first year box office receits in millions
totproducst	Numeric	8	0	total production costs in millions
totpromcst	Numeric	8	0	total promotional costs in millions
totbooksale	Numeric	8	0	total book sales in millions

The *dependent* (criterion) variable is the first year box office receipts in millions.
The *predictor* variables in this model are the total production costs in millions, total promotional costs in millions, and the total book sales in millions.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.983[a]	.967	.950	7.541	.967	58.221	3	6	.000

a. Predictors: (Constant), total book sales in millions, total promotional costs in millions, total production costs in millions

[18] This data set is part of the ***Technology Guide*** (ISBN: 0-618-20557-8) and ***Excel Guide*** (ISBN: 0-618-20556-X) that accompany, ***Understandable Statistics***, 7e. To access this data set go to http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets. Choose Data for **multiple linear** regression, and select ***Hollywood Movies***.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9932.463	3	3310.821	58.221	.000[b]
	Residual	341.201	6	56.867		
	Total	10273.664	9			

a. Dependent Variable: first year box office receits in millions

b. Predictors: (Constant), total book sales in millions, total promotional costs in millions, total production costs in millions

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7.676	6.760		1.135	.299
	total production costs in millions	3.662	1.118	.421	3.276	.017
	total promotional costs in millions	7.621	1.657	.559	4.598	.004
	total book sales in millions	.828	.539	.127	1.536	.175

1. State the null and the alternative hypotheses.

2. Evaluate the results and answer the following questions

- Is the alternative hypothesis supported? Which statistic confirms it?
- Given the results, which factors predict the first year sales of movie tickets and which do not? Which statistics provide us with this information?

Answers to Guided Practice Exercise

1. H_0: The regression equation (with total production costs in millions, total promotional costs in millions, and the total book sales in millions) does not significantly predict the variance in the first year movie sales, i.e., the first year box office receipts, Y scores.

   H_1: The regression equation (with total production costs in millions, total promotional costs in millions, and the total book sales in millions) significantly predicts the variance in the first year movie sales, i.e., the first year box office receipts, Y scores.

2. Evaluation of the results:

- Yes, F =58.22, and it is statistically significant
- Two factors predict the first year movie sales: production and promotional cost, but not the book sales. We derive this information from the coefficients table. The beta coefficients for production and promotional cost, but not for book sales, were statistically significant.

References

Cengage Learning. Textbook site for Understandable Statistics, Data sets. Retrieved on August, 9, 2015, from http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/dataset

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Morgan, G. A., Griego, O. V., & Gloeckner, G. W. (2001). *SPSS for Windows: An introduction to use and interpretation in Research.* Lawrence Erlbaum Associates, Publishers.

Chapter 10: Writing Like a Scientist

Introduction

Communication of scientific knowledge can be transmitted through several means, writing and publishing a manuscript in an academic journal, presenting it at a conference in the form of a poster, or by giving a talk. In this chapter we will focus on the first option—writing a manuscript. We will go over the structure of a typical academic paper, and also touch on some of its important stylistic features.

10.1 Know the Purpose of Your Paper

Before you begin the writing process, make sure that you are clear about what the purpose of your paper. This may sound strange, but after reading many undergraduate students' research papers, I have come to realize that students don't always know what they are trying to communicate—they lack a, "take home message." And, if the writer is not focused, or struggling with the message, it is almost certain that the reader will be confused.

Broadly speaking, you will be either, reporting results of your own original study or, you will be reviewing and evaluating empirical and theoretical articles written by other people. While these are different academic papers, the general purpose of both is to share new information, e.g., novel ideas and/or findings of a study. However, it is not enough to have an interesting idea. The research has to be communicated in such a way that the audience can appreciate it as well. Two things will help you accomplish this task: adhering to a format that is accepted in a field of your discipline, and writing in academic style.

Facts, Claims of Facts, and Opinions

Perhaps the main difference between an academic or a scholarly writing from any other writing is that the former must rely on facts and evidence; and cannot express personal opinions. Although this sounds as no big surprise and even easy to do, most students will find it quite challenging; starting with being able to distinguish between facts, claims of facts and personal opinions. Typically, in an academic paper, a fact is something that we, as a society, have agreed upon to be true based on prior overwhelming support of scientific evidence.

Examples of some of such facts are that the earth is round and evolves around the sun; the fact that as we mature we become cognitively more advanced and intelligent; that children rely on adults

to take care of them physically and emotionally; the fact that people seek pleasure and dislike unpleasant stimuli; people vary in their opinions and preferences as much as they vary in their physical appearances. However, more often when we write, we make claims of facts, in which case, we must substantiate them with evidence, by citing scientific articles.

So, what about findings of well-designed and executed scientific studies? Can they be considered facts? The answer is no. No one scientific study is good enough to yield 100% facts.

There are at least four good reasons, that I can think of, for that:

(1) Most all empirical investigations are based only on a sample not an entire population of interest; thus, any inferences will have some level of uncertainty (perhaps the sample wasn't big enough or representative enough).

(2) No one study is capable of controlling for or including all relevant to the phenomenon variables to account for the found effect (or lack thereof).

(3) Even if a study used an entire population and included all relevant variables (which is virtually impossible), it may not be free of errors (e.g., in execution, calculations, etc.).

Finally, (4) most results that come from scientific investigations are not always straightforward; they still have to be interpreted, in which case, there may be several explanations and interpretations of the same data.

Therefore, as you should have already learned, the best remedy for any errors or weaknesses of a scientific investigation, and the best way to ensure that the claims are in fact true is through replication; the more studies and of different kinds find the similar results, the more likely that the claims are facts. In short, your best bet is to treat any scientific finding as a claim of fact and support it with evidence.

What about opinions? Personal opinions should be excluded; after all, the whole purpose of an academic paper is to advance true knowledge through facts and scientific evidence (Recall that the most reliable, free of bias, information comes from science). You can, however, express an 'educated guess' or make a speculative statement if it is grounded in some evidence. In other words, if the amount of evidence is weak or the evidence is only indirect, you can hypothesize as to the true nature of the found results. However, you have to let your reader know that the claims are only speculative and must be treated with a healthy amount of skepticism.

Proper Language and Citations

When you write, you have to keep the distinction between facts, claims of facts and opinions in mind for at least two reasons.

(1) As I mentioned in the above section, facts are mutually agreed upon truths, and don't require convincing your reader. Claims of facts, on the other hand, need to be substantiated with evidence. (Opinions should be excluded all together). Thus, when you make a claim, you must cite some evidence—a scientific article or multiple articles (even better) supporting your statement. If you don't cite your claims, your reader will be unable to take your claims seriously and your efforts in persuading your reader will be wasted.

(2) Another important reason to keep these distinctions is that the language that you use to make statements of facts or claims will be (and should be) different. If something is a matter of fact, you state it with certainty. For example, I can state with certainty, that there are some biological differences between men and women—that is a matter of fact. However, to further claim that some of these biological differences cause differences in men's and women's behavior is a claim of fact that will require citing scientific. Furthermore, my language should convey some level of uncertainty, for example, through the use of such words as 'possibly, 'potential', 'likely', etc.

Here are a couple of examples illustrating a sentence of a matter of fact and a claim of fact:

There are certain biological differences between men and women, for example, women are physically capable to bear children while men aren't (This is a matter of fact). *Childbearing **may have** economic consequences for women, such as women having lower wages than men and/or women occupying lower level positions* (e.g., Waldfogel, 1997) (This is a claim of fact and so I added a citation of an empirical study that found a 6% drop in women's wages after having one child).

Clear and Simple

Contrary to what many students might think, academic writing should be clear and simple. You should not purposefully complicate your sentences or substitute simple words with fancy ones just to sound intelligent. Doing so will make your writing confusing, pompous, and you will probably make more grammatical mistakes.

- **Example:** do not substitute *'large'* for *'abundant' [effect]*

Concise

Every sentence should be meaningful, don't waste your space with filler sentences or vague/empty phrases. Make every sentence count.

- **Example:** ***Considering that we have now established that the media does portray stereotypes, next we should delve into what kind of stereotypes are to be expected. To start,*** *beautiful people are primarily viewed to be a cut above the rest in every field, they have an undeserved veneration (Anderson et al. 2008).*

The words in bold seemed too wordy and unnecessary. The writer could have simply omitted the entire highlighted sentence and the phrase 'to start' to begin describing various kinds of stereotypes on TV.

Objective and Inclusive

Since you are writing a scholarly paper, you have to choose your words wisely to stay objective and factual. According to APA bias can be defined as partiality or " an inclination or predisposition for or against something" (see https://www.apa.org/about/apa/equity-diversity-inclusion/language-guidelines.pdf (p.2). Do not choose words or phrases that can be biased, or sound judgmental or emotional, and don't overstate or over-generalize findings (Rowena, 2011, p. 221).

- **Example 1:** *"This can ruin the hopes and dreams."*— This type of statement is emotional and over-generalized. It can be re-stated as "This can negatively impact the attitudes or the self-perception which can in turn influence the motivation to succeed..." (this is a more precise and neutral in tone).
- **Example 2**: *"There is not a shadow of doubt that cosmetics are truly effective at improving social perceptions that the person who's wearing it may wish to gain."*—There is always room for doubting in science, in fact, it is the bedrock of scientific explorations. It can be restated as "It is likely that cosmetics effect how the person wearing it is perceived."

The language must also be inclusive--represent all possible social and cultural groups and identities. Please see APA inclusive language guidelines for more details:

https://www.apa.org/about/apa/equity-diversity-inclusion/language-guidelines.pdf

Formal

Perhaps, the most obvious recommendation is to be formal; don't write the way you speak. Conversational language can be vague, inaccurate, offensive/judgmental, emotional, biased, or all of the above.

- **Example:** instead of *"grown women"* write *"adult women"*

Perspective and Voice

APA style now encourages to use first person voice when describing the research process, which is normally written in the Method section of the paper. Therefore, you should use 'I' when referring to what you did and the steps of the research process you undertook. Do not use 'we' if you are the only researcher who conducted the study. You can also write in a third voice perspective because it helps give the appearance of objectivity. Use active voice (as much as possible) as it conveys information with more clarity and precision. For example, instead of writing that ***participants were interviewed***, you can restate it using active voice to clarify who tested them, e.g., a ***research assistant interviewed the participants***. Finally, it is best to highlight the research rather than the researcher. For example, instead of writing that "researchers found…" you should state that "the study found…" something. Again, this gives more objectivity to your words.

Font and Size

APA allows several fonts. But make sure that whatever font you choose to use, first check with your instructor and the requirements for the paper in your course. Second, stick to the same font and size throughout the paper. Finally, when in doubt, (and this is my personal recommendation) use 12-point Times New Roman.

Miscellaneous Issues: Margins, Line Spacing, Alignment, etc...

You should use 1-inch margin for every page of your paper. Almost all parts of your paper should be double-spaced. The exceptions are for the title page (see title page examples), tables, figures and footnotes. Alignment should be to the left. A paragraph should be indented 0.5 in from the left margin (You can simply use the tab to do that).

Tense

APA recommends using the following verb tenses:

When you are reviewing previous studies and citing articles, use ***past or present perfect tense***.

For example: Hefner and Wilson (2013) found no differences in romantic expressions (past tense). Many studies have found gender differences in prosocial behavior (present perfect tense).

When you are describing the procedures use ***past tense***.

For example: The subjects completed a survey (past tense).

When you are reporting results of your study, use ***past tense***.

For example: The test revealed a significant difference between the two groups.

When you are discussing the findings and making general statements, use present tense.

For example: The findings of the study suggest that females do not hold idealistic beliefs.

10.3 Format Style

Psychology and many other disciplines have adopted an academic writing style of the American Psychological Association (APA). As a student of psychology and the social sciences, you will have to follow this style when you write your research paper. In this chapter I will go over only the basics of an APA format, that is, the structure and the citations. The complete coverage of APA style is described in the Publication Manual of the American Psychological Association, 6th and 7th editions. There are some small differences between the 6th and 7th editions. One main difference is the formatting of the title page. The 7th edition includes specific instructions for the title page of a student paper. Please see examples later in the chapter. Another difference is that you don't have to have a running head if you are writing a student paper, according to the 7th edition.

The following are the main sections of an APA-style manuscript:

Title Page— *separate page (section 1)*

Abstract— *separate page (section 2)*

Introduction *(section 3)*

Methods *(section 4)*

Results *(section 5)*

Discussion *(section 6)*

This is the main body of your paper. Each section should flow naturally from the previous one—do not start one on a separate page.

References—*separate page (section 7)*

Appendices—*separate pages (section 8)*

Each section will be discussed next.

Section 1: Title Page

The title of your paper goes in the center, followed by your name, and the name of your Institution. The Author note is mainly for published articles. This should include the author's name, institutional affiliation, and any changes in his/her affiliations.

The title of your paper goes in the center, followed by your name, and the name of your Institution. The Author note is mainly for published articles. This should include the author's name, institutional affiliation, and any changes in his/her affiliations.

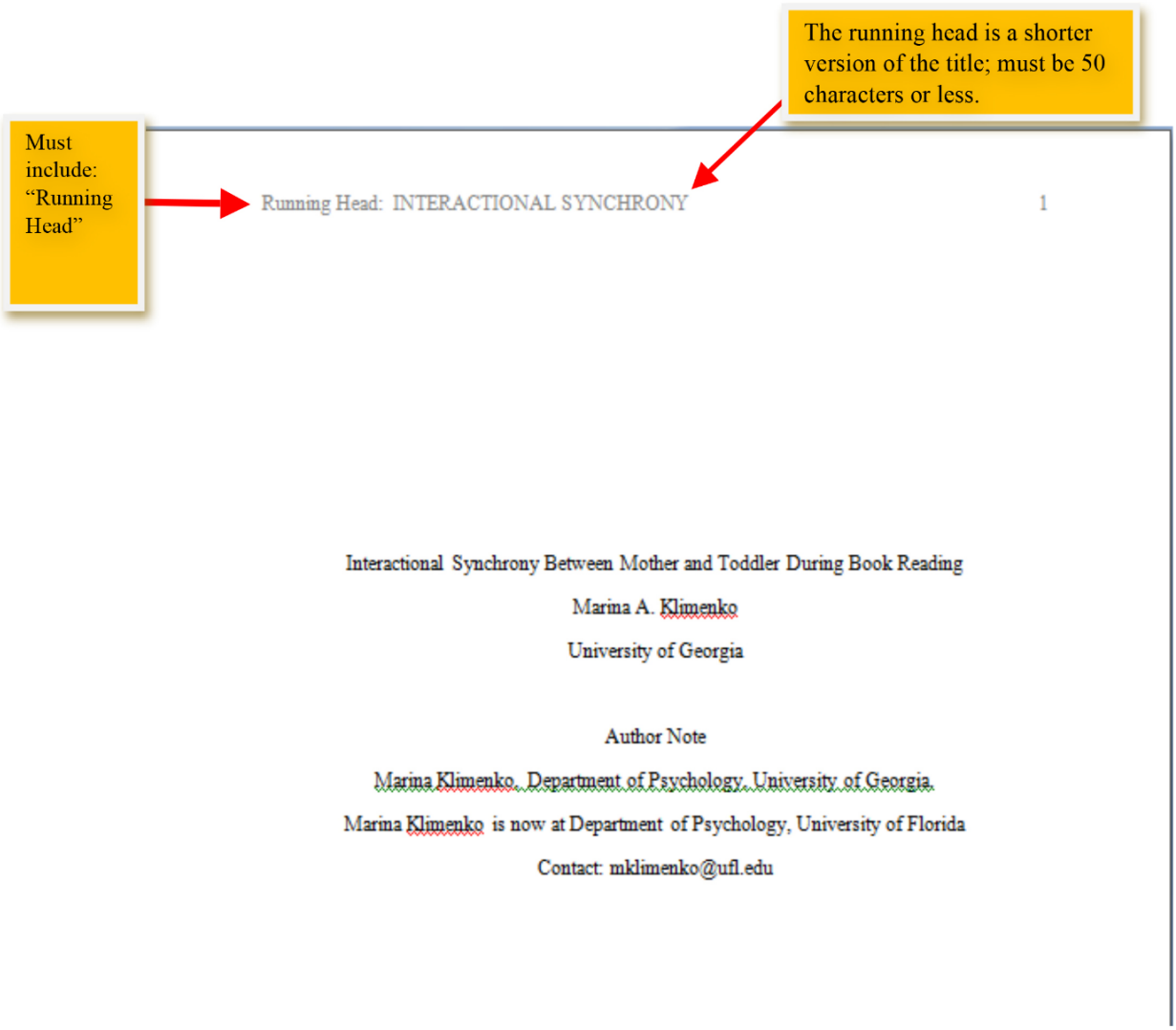


Figure 10.1. *Sample of a Title Page. Paper adapted from Klimenko, M. (2007). Interactional synchrony between mother and toddler during book reading (Master's thesis).*

According to the 7th edition rules, student papers do not have to have a running head. Please see example below of how to format a title page if you are writing a student paper:

Consequences of Romantic Expressions in Film among Gender and Targeted Age Groups

First Name Last Name

Department of Psychology

PSY 3213L: Research Methods Lab, University of Florida

Dr. Marina A. Klimenko

March 18, 2021

The title page of a student paper should include the page number as the only page header, student's name, school affiliation, name of the course, instructor's name, and the due date of the assignment.

Section 2: Abstract

Provide a brief summary of your manuscript (between 150 and 250 words): purpose/methods, results, and "take home message" (see Figure 10.2).

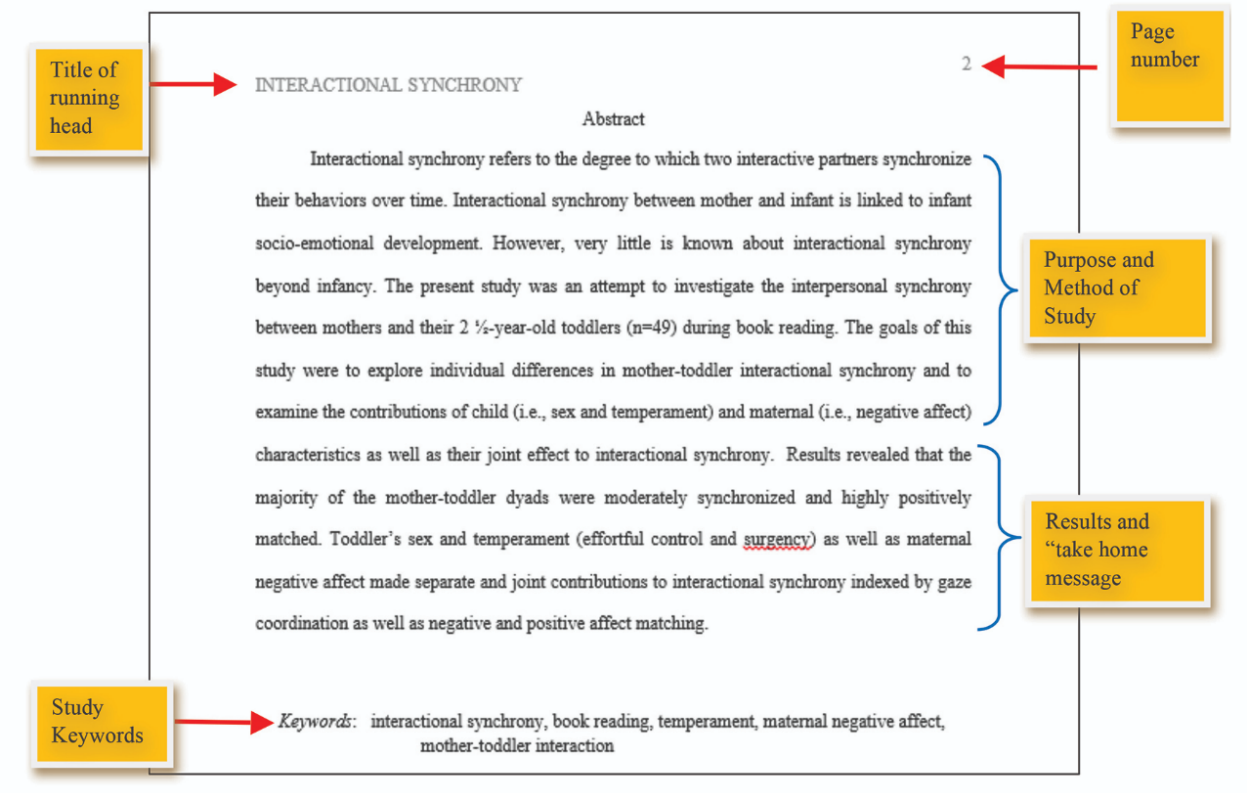


Figure 10.2. An abstract from Klimenko, M. (2007). Interactional synchrony between mother and toddler during book reading (Master's thesis).

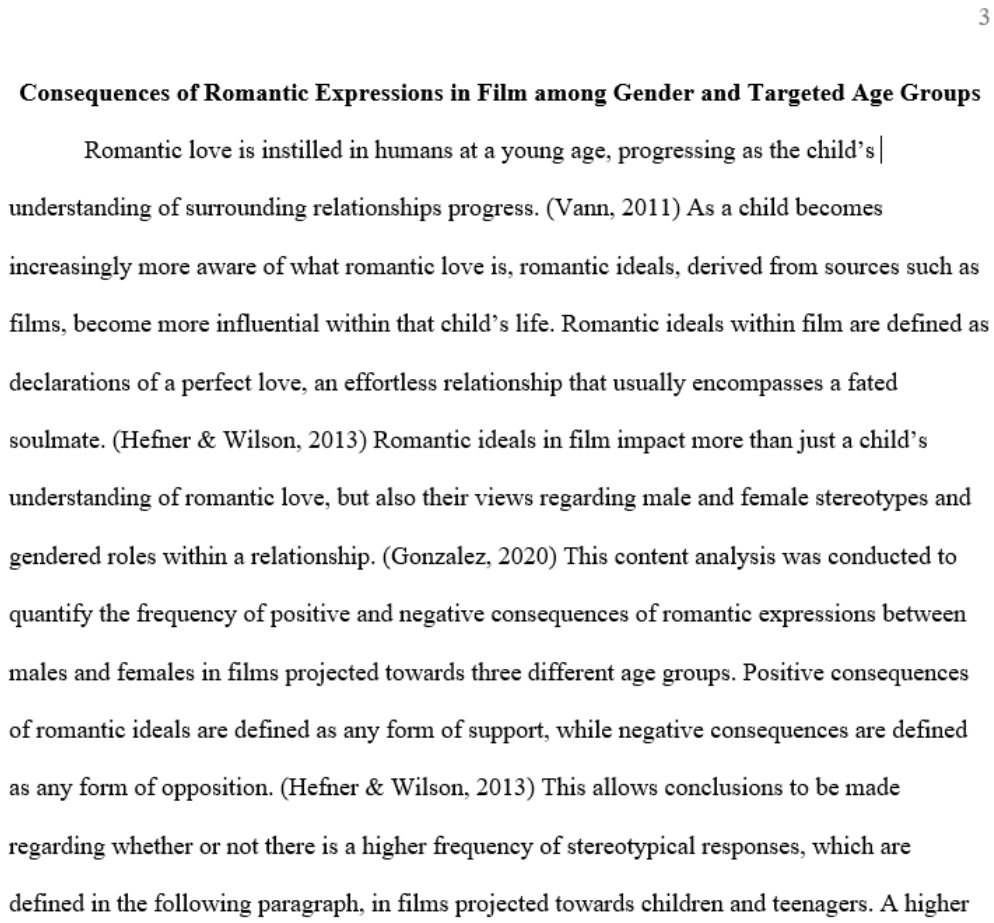
Main Body: Format and Levels of Heading (7th Edition)

Within the main body of your text, you will have different sections (e.g., introduction, methods, results, etc.). Furthermore, each section may have one or more sub-sections. Therefore, to identify each section and sub-section (section within the main section), you need to use headings. Levels of headings will help the reader identify different sections/sub-sections. There can be a maximum of 5 levels (or less). Each level will have to have a different style of heading. Please see the examples of the levels and their heading style:

Level	Format	Examples
1	**Centered, Bold, Capitalized**	**Title of the Paper, Methods, Results**
2	Flush Left, Bold, Capitalized	**Procedures**
3	*Flush Left, Bold Italic, Capitalized*	***Measures***
4	Indented, Bold, Capitalized, Ending With a Period.	**Independent Variables.**
5	*Indented, Bold Italic, Capitalized, Ending With a Period.*	***Operational Definitions.***

Main Body: Text

The main body of your paper begins with an “Introduction” section. You should put the actual title of your paper, and not "Introduction" at the very top of the text, and centered (see Figure 10.3). Notice also that it should be in bold (Level 1 heading).



3

Consequences of Romantic Expressions in Film among Gender and Targeted Age Groups

Romantic love is instilled in humans at a young age, progressing as the child's understanding of surrounding relationships progress. (Vann, 2011) As a child becomes increasingly more aware of what romantic love is, romantic ideals, derived from sources such as films, become more influential within that child's life. Romantic ideals within film are defined as declarations of a perfect love, an effortless relationship that usually encompasses a fated soulmate. (Hefner & Wilson, 2013) Romantic ideals in film impact more than just a child's understanding of romantic love, but also their views regarding male and female stereotypes and gendered roles within a relationship. (Gonzalez, 2020) This content analysis was conducted to quantify the frequency of positive and negative consequences of romantic expressions between males and females in films projected towards three different age groups. Positive consequences of romantic ideals are defined as any form of support, while negative consequences are defined as any form of opposition. (Hefner & Wilson, 2013) This allows conclusions to be made regarding whether or not there is a higher frequency of stereotypical responses, which are defined in the following paragraph, in films projected towards children and teenagers. A higher

Figure 10.3. *Sample student paper.*

The purpose of the *introduction* is to create interest in your paper and to set the stage for what will be coming in the rest of the paper, review of the literature (if it is a review paper, in some cases, an empirical paper can also have a separate literature review sub-section) or description of the study conducted by the author. To this end, you, the author, need to explain the purpose and the rationale of your study, by reviewing previous literature and identifying any gaps in the current knowledge. Your hypothesis will need theoretical and empirical justification as well.

This is the major part of your paper, and so, you need to plan the logical flow of this section. I like the idea of constructing a mental scheme of what you are going to write and in what order (see Zhang, 2014). But generally speaking, you can sequence it as follows: Begin by introducing the topic of your investigation and highlighting its importance. A more detailed review of the literature should follow, demonstrating what is known and what questions still remain unanswered (see Figure 10.4). Reviewing literature doesn't mean that you have to simply summarize every article

that you can find. Focus instead on the articles and the details that are relevant only to your paper, otherwise you will create confusion and disrupt the flow of your logic.

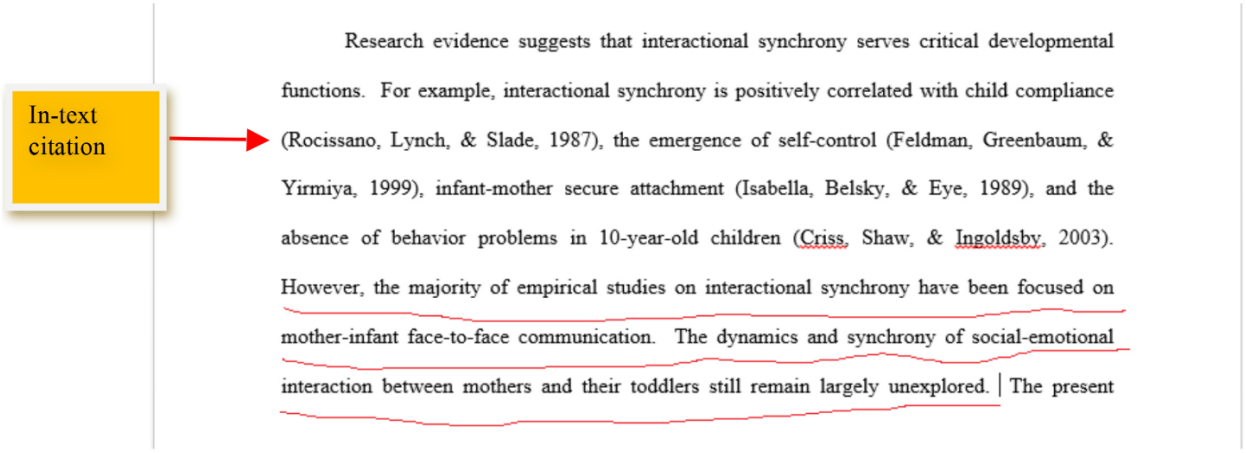


Figure 10.4. *Example of a Literature Critique/Gap; adapted from Klimenko, M. (2007). Interactional synchrony between mother and toddler during book reading (Master's thesis).*

Cite the articles that you are reviewing or using as evidence by listing only the last names of the author(s) and the year of the publications (see Figure 10.4). If there are more than 6 authors, or, if this is the second time you are mentioning the same study, list only the first author, followed by [et al., 2001] (see Figure 10.5). When you refer to an author's name or a specific publication, write their last name only, followed by the year of the publication, in parentheses (see Figure 10.5).

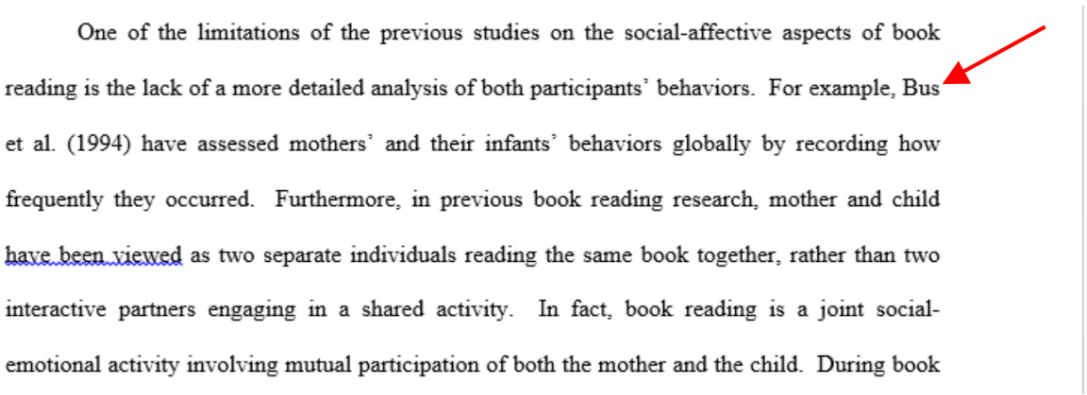


Figure 10.5. *Example of an In-Text Citation; adapted from Klimenko, M. (2007). Interactional synchrony between mother and toddler during book reading (Master's thesis).*

Finally, conclude the section by stating your hypothesis and providing a brief synopsis of how it *was* addressed. Make sure that any reference to your own study is done in the past tense. For example, instead of writing that your hypothesis "***will*** be addressed," it should be written as "***was*** addressed" (see Figure 10.6).

In sum, the present study was aimed to investigate individual differences in interactional synchrony among mother-toddler dyads during book reading and to examine the separate as well as joint contribution of child (i.e., sex and temperament) and maternal (i.e., negative affect) characteristics to book-reading interactional synchrony.

Figure 10.6. *Conclusive Statements; adapted from Klimenko, M. (2007). Interactional synchrony between mother and toddler during book reading (Master's thesis).*

Final advice: Refer to your study as a *study*, *research*, or an *investigation*. IF the method of your study was experimental then you can call it an *experiment*. A study is not automatically an experiment just because it is scientific.

Main Body: Method

This section begins with the heading "**Method**", in bold and centered (see Figure 10.7).

The purpose of this section is to describe the process, and all the relevant materials, that you had to use to collect your data. Provide enough detail so that the reader can judge the soundness of your research design and replicate your study. To make it easier to follow this section, it can be further broken down into "Participants", "Apparatus and materials", "Procedures", and/or "Data analysis".

Participants

Describe your participants (people) or your subjects (animals), how many were enrolled in the study, their age, gender and racial/ethnic composition (for human participants) (see Figure 10.7). However, be mindful that according to the ethical rules of APA Publication Manual, confidential

or any information that can personally identify research participants cannot be revealed. Finally, in some research (e.g., content analysis), no participants or subjects are used, which means that this subsection should be omitted.

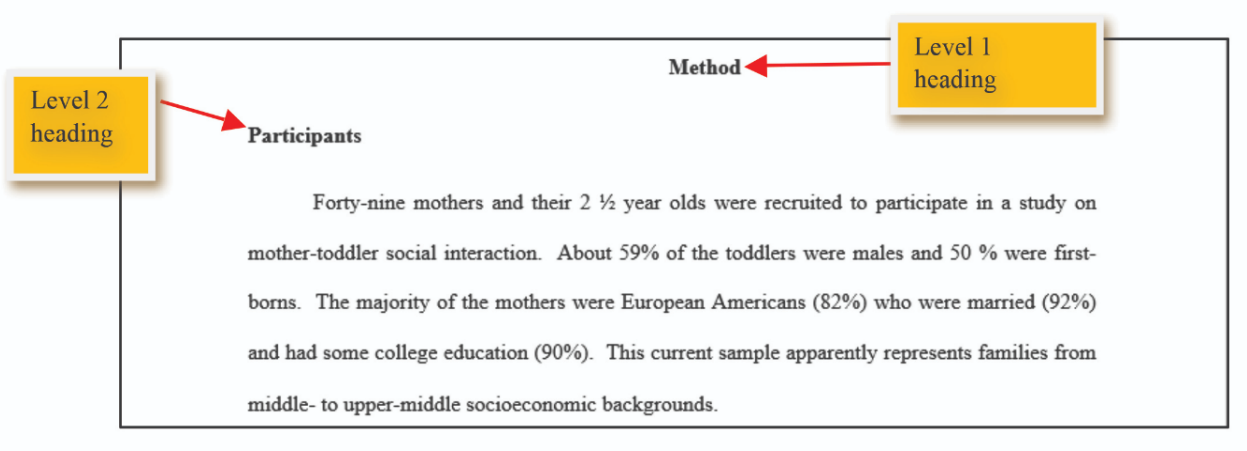
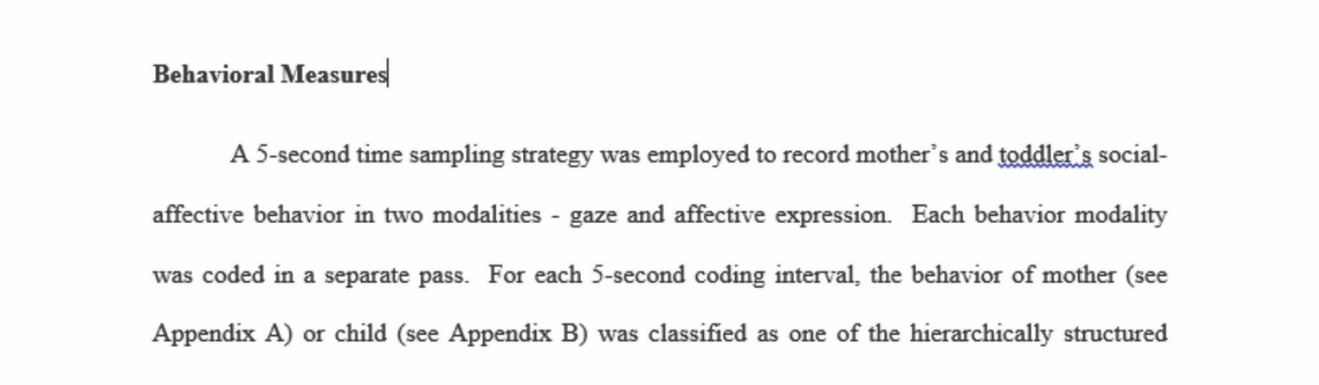


Figure 10.7. *Example of Method (Participants) Sub-section; adapted from Klimenko, M. (2007). Interactional synchrony between mother and toddler during book reading (Master's thesis).*

Apparatus and materials (or behavioral/other measures).

Describe any equipment or apparatus used in the study. This is where you should also describe any behavioral or other psychological measures (see Figure 10.8).

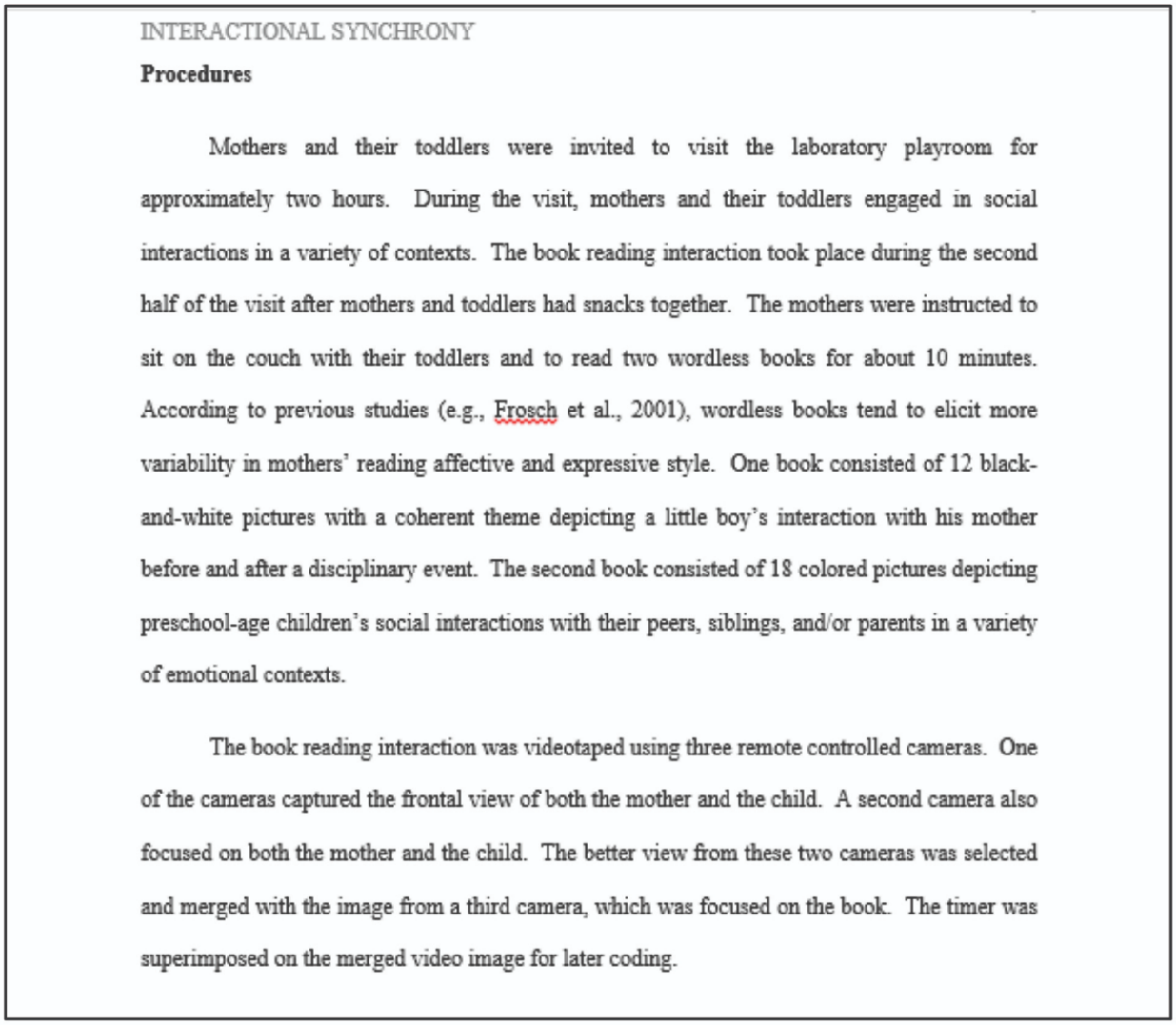

Behavioral Measures

A 5-second time sampling strategy was employed to record mother's and toddler's social-affective behavior in two modalities - gaze and affective expression. Each behavior modality was coded in a separate pass. For each 5-second coding interval, the behavior of mother (see Appendix A) or child (see Appendix B) was classified as one of the hierarchically structured

Figure 10.8. *Example of Method (Behavioral Measures) Sub-section; adapted from Klimenko, M. (2007). Interactional synchrony between mother and toddler during book reading (Master's thesis).*

Procedures

Describe the process that was used to collect the data (see Figure 10.9).



INTERACTIONAL SYNCHRONY

Procedures

Mothers and their toddlers were invited to visit the laboratory playroom for approximately two hours. During the visit, mothers and their toddlers engaged in social interactions in a variety of contexts. The book reading interaction took place during the second half of the visit after mothers and toddlers had snacks together. The mothers were instructed to sit on the couch with their toddlers and to read two wordless books for about 10 minutes. According to previous studies (e.g., Frosch et al., 2001), wordless books tend to elicit more variability in mothers' reading affective and expressive style. One book consisted of 12 black-and-white pictures with a coherent theme depicting a little boy's interaction with his mother before and after a disciplinary event. The second book consisted of 18 colored pictures depicting preschool-age children's social interactions with their peers, siblings, and/or parents in a variety of emotional contexts.

The book reading interaction was videotaped using three remote controlled cameras. One of the cameras captured the frontal view of both the mother and the child. A second camera also focused on both the mother and the child. The better view from these two cameras was selected and merged with the image from a third camera, which was focused on the book. The timer was superimposed on the merged video image for later coding.

Figure 10.9. *Example of Method (Procedures) Sub-section; adapted from Klimenko, M. (2007). Interactional synchrony between mother and toddler during book reading (Master's thesis).*

Data Analysis

This subsection is reserved for the description of any statistical analyses the author had to perform to complete the study. In some cases, researchers omit this subsection and include the description of their statistical analyses as they report the results in the *results* section.

It is also a good idea to check with your course instructor, or with the publication instructions of a journal you are planning to submit your manuscript to, to see whether any additional details or subsections have to be included or omitted.

Main Body: Results

The heading "**Results**" must be in bold, and centered (level 1 heading; see Figure 10.10). This is the place for reporting the results of your descriptive and inferential statistical analyses which you carried out to address the goals of a study. To make the content of this section more readable, organize it and report the findings in the same order as your research questions or your hypotheses as listed in the introduction. But, start with the descriptive statistics first; summarize the sample and the variables of your study. This information is important when the reader is trying to understand the results of the inferential statistical tests.

To make this part of your manuscript readable and concise, describe your findings in a narrative form using tables and graphs. Choose the form that is most effective and efficient. For example, if there are only two means, report them in written form rather than creating a separate table (see Table 10.10). On the other hand, if you have to provide means and standard deviations for multiple variables, it would make more sense to provide that information in a table. Similarly, use graphs only if they help your reader understand your story line rather than to show off your graphing skills. Graphs should add new information rather than repeat what was already described. Your tables and figures have to go to Appendices, the last section in a manuscript (see Figure 10.17).

Finally, remember that your hypothesis was either *supported* or not *supported*, never proved. No study can ever give us answers with 100% certainty.

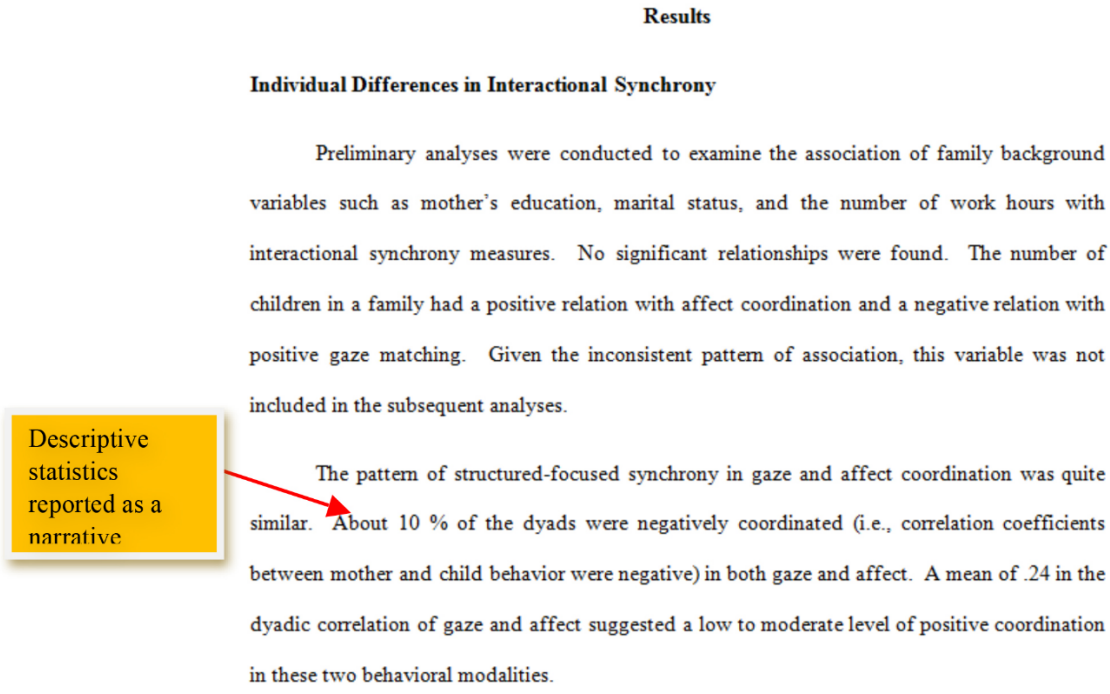


Figure 10.10. *Example of Results Sub-section; adapted from Klimenko, M. (2007). Interactional synchrony between mother and toddler during book reading (Master's thesis).*

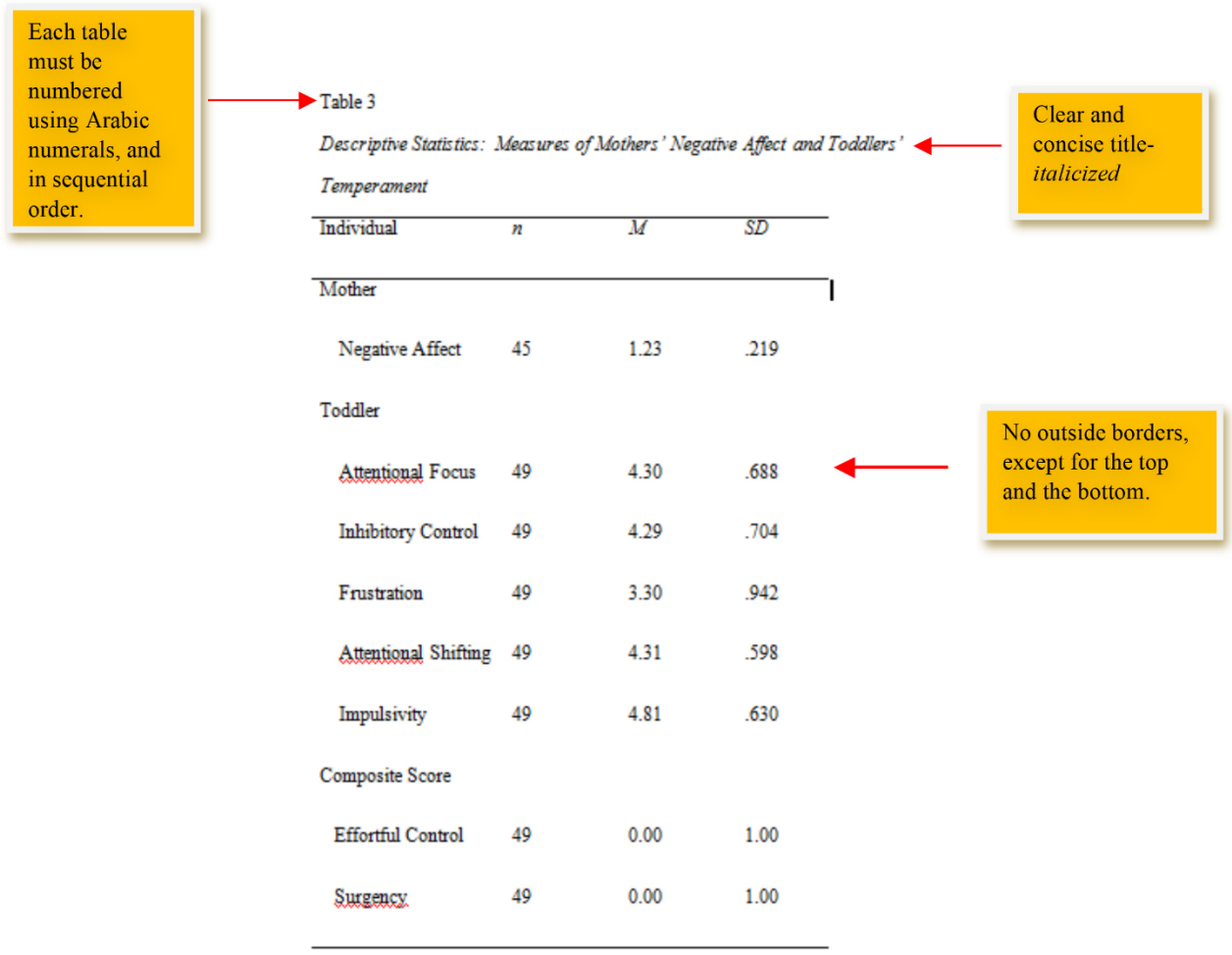
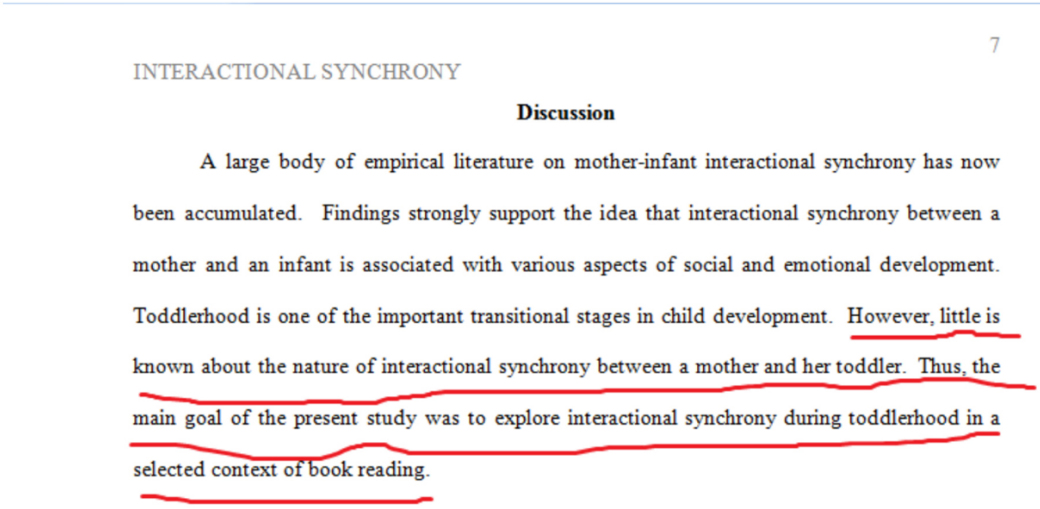


Figure 10.11. *Example of a Table in APA format; adapted from Klimenko, M. (2007). Interactional synchrony between mother and toddler during book reading (Master's thesis).*

Main Body: Discussion

The heading, "Discussion" (level 1 heading) has to precede the text, in bold and centered (see Figure 10.12). Results are not conclusion of the study. After reporting the results of your statistical analyses, you will have to **interpret** them in this section by going back to your original research questions and hypotheses. You will have to explain what the general rule of the relationship between the variables can be drawn. This will entail addressing the following points and, typically, in the following order:

- Reminding your reader what the purpose of your research was by restating your hypotheses and/or listing your research questions (see Figure 10.12)



7

INTERACTIONAL SYNCHRONY

Discussion

A large body of empirical literature on mother-infant interactional synchrony has now been accumulated. Findings strongly support the idea that interactional synchrony between a mother and an infant is associated with various aspects of social and emotional development. Toddlerhood is one of the important transitional stages in child development. However, little is known about the nature of interactional synchrony between a mother and her toddler. Thus, the main goal of the present study was to explore interactional synchrony during toddlerhood in a selected context of book reading.

Figure 10.12. *Example of a Discussion (restating goals/hypotheses); adapted from Klimenko, M. (2007). Interactional synchrony between mother and toddler during book reading (Master's thesis).*

- Explicitly stating whether support to your hypotheses or your predictions have been found. State if the findings were not what you expected, try to find an explanation. If there is more than one hypothesis/ question to be discussed, create additional level 2 (and 3) headings, and discuss them in the order they were introduced and tested (see Figure 10.13)

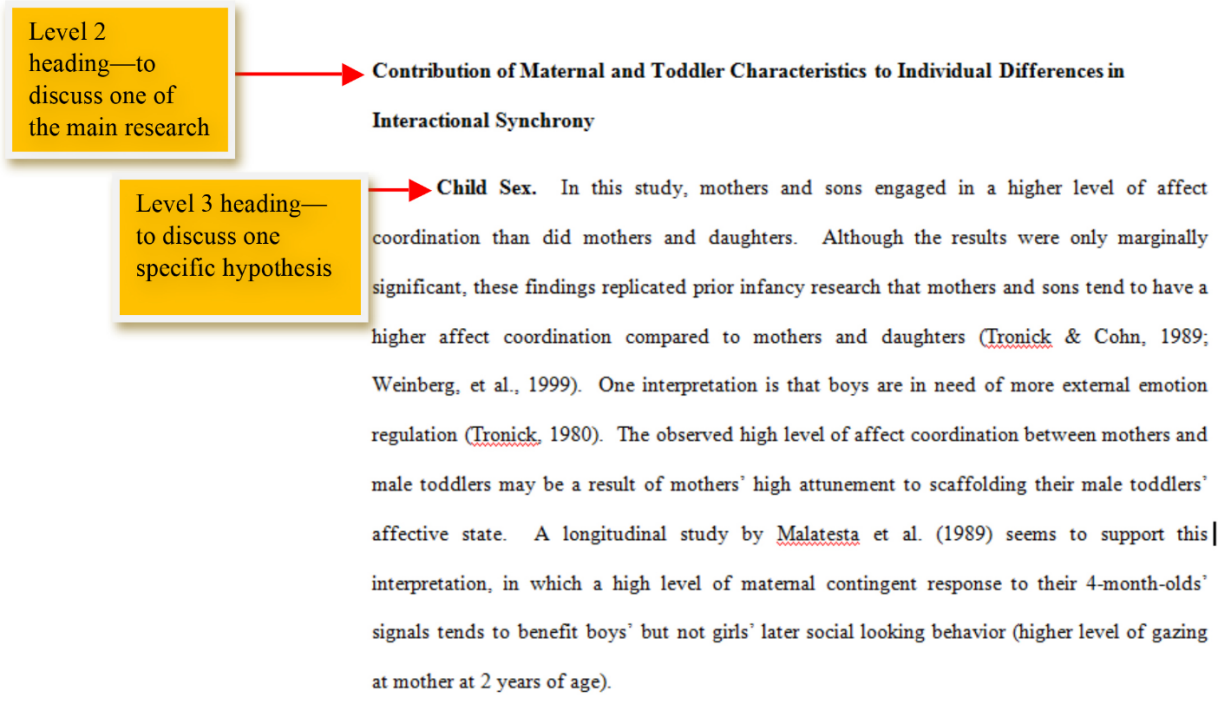


Figure 10.13. *Example of a Discussion (interpretation of results); adapted from Klimenko, M. (2007). Interactional synchrony between mother and toddler during book reading (Master's thesis).*

- Discussing any possible limitations of your study (they can be placed under one subheading with future directions as in my example; see Figure 10.14)

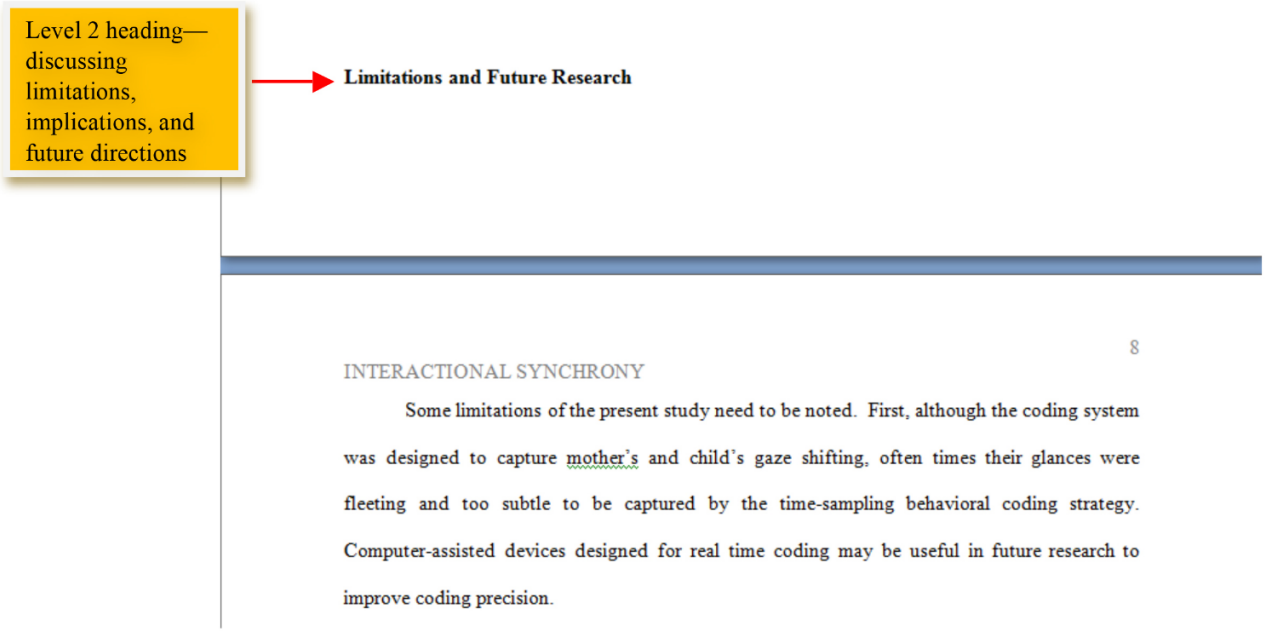


Figure 10.14. *Example of a Discussion (limitations, implications and future research directions); adapted from Klimenko, M. (2007). Interactional synchrony between mother and toddler during book reading (Master's thesis).*

- Explaining how your findings can advance knowledge in the field and giving suggestions for future research directions (see Figure 10.15). No single study should be viewed in isolation. Therefore, help your reader understand how the findings of your study fit within the broader body of current evidence.

Finally, regardless of how interactional synchrony is operationalized, this dyadic construct can potentially be a fruitful research area for our understanding of the precursors of development in early childhood. For example, taking a social-constructive approach to social cognition, Carpendale and Lewis (2006) proposed that young children develop social understanding through cooperative interactions with others. Mother-toddler synchrony may present a context for young children to develop an understanding of other people's intentions, feelings, beliefs, and desires.

Figure 10.15. *Example of a Discussion (advancing knowledge and future research directions); adapted from Klimenko, M. (2007). Interactional synchrony between mother and toddler during book reading (Master's thesis).*

Section 7: References

The references section must begin on a separate page. Put "References" at the top of the page, in bold and centered (level 1 heading). All cited sources must be listed here, by including their full references: authors' names (in alphabetical order), title/year of the publication/book (pages and publisher's name).

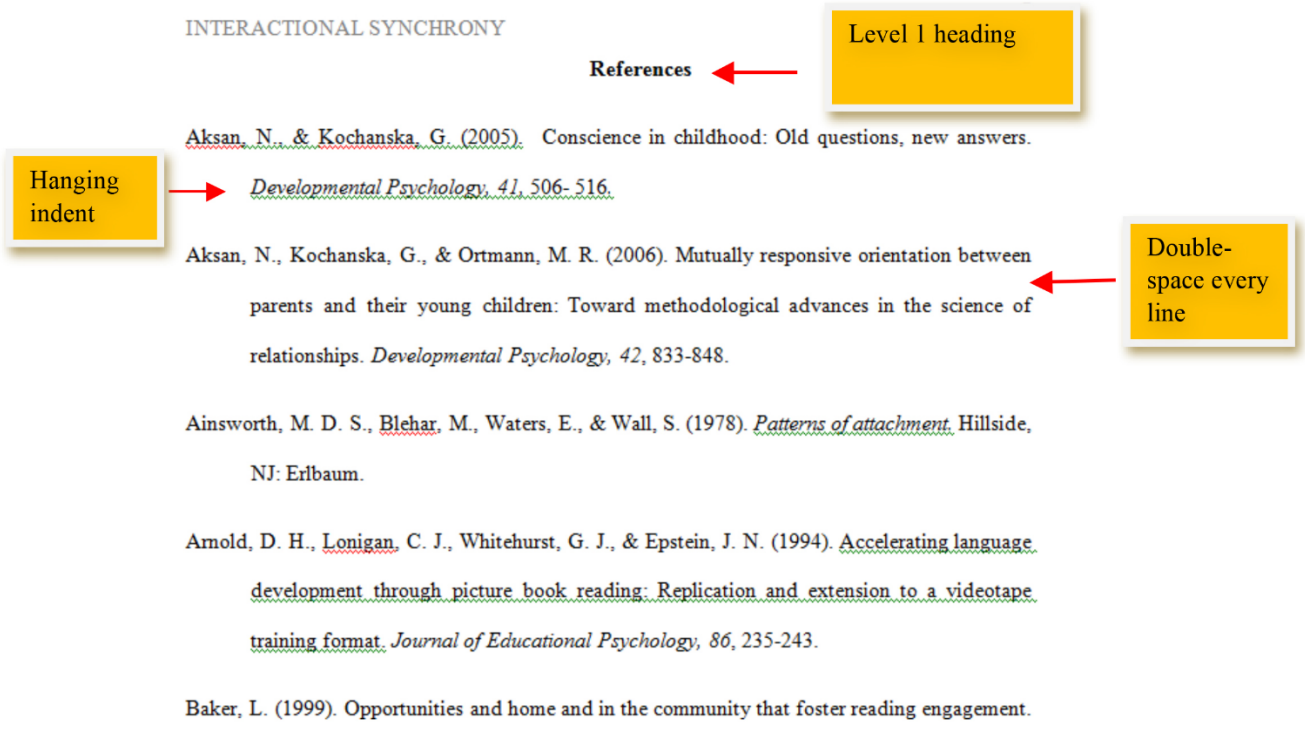


Figure 10.16. *References section; adapted from Klimenko, M. (2007). Interactional synchrony between mother and toddler during book reading (Master's thesis).*

One Author. Cite the work with one author as follows:

Chapman, M. (1979). Listening to reason: Children's attentiveness and parental discipline. *Merrill-Palmer Quarterly*, 25, 251-263.

Two Authors. A work with two authors should be cited as follows:

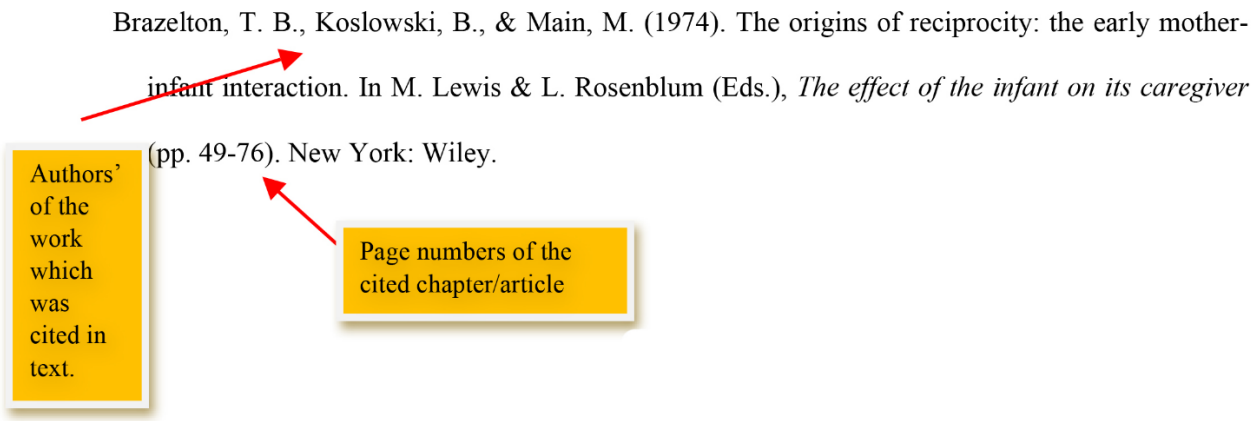
Aksan, N., & Kochanska, G. (2005). Conscience in childhood: Old questions, new answers. *Developmental Psychology, 41,* 506- 516.

Three to Six Authors. The following is an example of a cited work with, more than two, butless than seven authors:

Bus, A. G., Van Ijzendoorn, M. H., & Pellegrini, A. D. (1995). Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy. *Review of Educational Research, 65,* 1-21.

Citation of a Book. If the work to be cited is a book, use the following format:

Ainsworth, M. D. S., Blehar, M., Waters, E., & Wall, S. (1978). *Patterns of attachment*. Hillside, NJ: Erlbaum.



Online Press Release. If you want to cite an online press release, format it as follows:

American Psychological Association. (2021, July 12). Officers' tone of voice reflects racial disparities in policing [Press release]. https://www.apa.org/news/press/releases/2021/07/officers-tone-racial-disparities

To cite this in text, in parentheses, do as follows: (American Psychological Association, 2021) To mention it in writing in narrative form, do as follows: American Psychological Association (2021).

Tweet, Facebook, etc. If you want to cite an image or a post from a social media platform like Tweeter or Facebook, format it as follows:

American Psychological Association [@APA_style]. (2021, July 14). See the sample papers in Chapter 2 for examples of effective headings. [Image attached] [Tweet]. Tweeter. https://twitter.com/APA_Style

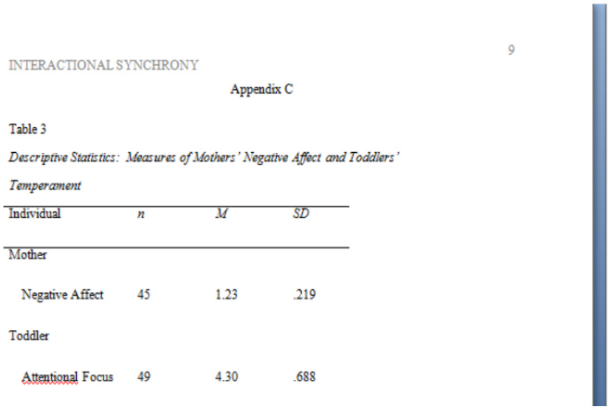
To cite it in text, in parentheses, do as follows: (American Psychological Association, 2021) To mention it in writing in narrative form, do as follows: American Psychological Association (2021).

These are just the most basic rules that you will most likely need to properly cite and format the reference page of your manuscript.

Section 8: Appendices

Place all your tables, graphs, and/or pictures in the appendices. Each item (e.g., table, graph) has to have its own appendix. Label each appendix as "Appendix A", "Appendix B", etc (see Table 10.17)

Place all your tables, graphs, and/or pictures in the appendices. Each item (e.g., table, graph) has to have its own appendix. Label each appendix as "Appendix A", "Appendix B", etc. (see Table 10.17)



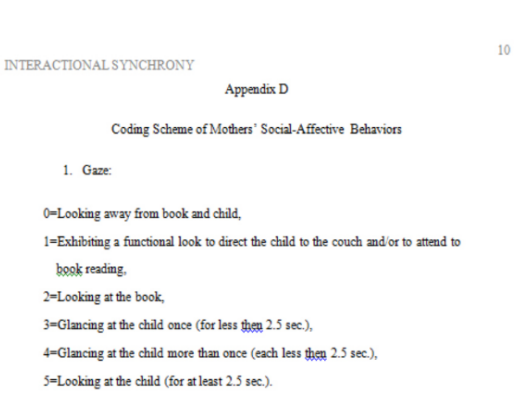
INTERACTIONAL SYNCHRONY 9

Appendix C

Table 3

Descriptive Statistics: Measures of Mothers' Negative Affect and Toddlers' Temperament

Individual	*n*	*M*	*SD*
Mother			
Negative Affect	45	1.23	.219
Toddler			
Attentional Focus	49	4.30	.688



INTERACTIONAL SYNCHRONY 10

Appendix D

Coding Scheme of Mothers' Social-Affective Behaviors

1. Gaze:

0=Looking away from book and child,
1=Exhibiting a functional look to direct the child to the couch and/or to attend to book reading,
2=Looking at the book,
3=Glancing at the child once (for less then 2.5 sec.),
4=Glancing at the child more than once (each less then 2.5 sec.),
5=Looking at the child (for at least 2.5 sec.).

References

Murray, R. (2011). *How to write a thesis* (3rd ed.). Maidenhead Berkshire, England: McGraw-Hill Education (UK).

Zhang, W. (2014). Ten simple rules for writing research papers. *Plos Computational Biology, 10*, 1-3. doi:10.1371/journal.pcbi.1003453

Chapter 11: Becoming a Critical Thinker and a Smart Consumer of Science

Introduction

This chapter was co-written with Dr. Ryan Mears.

Now that you know the basic theoretical underpinnings of science and have a fairly good understanding of the scientific process, let's talk about how this knowledge can help you be a more critical thinker and a smart consumer of science.

11.1 Healthy Skepticism

At the core of critical thinking lies healthy skepticism. We can go back to the ancient Greek philosophy (again), and trace the origin of this idea back to two main skeptical traditions, the Pyrrhonian and the Academics, of Hellenistic era and the Imperial age (Machuca, 2011). The Greek word skepsis means, literary, investigation, which should already tell you that the purpose of skepticism is to look for truth through doubt and inquiry. This can be further elaborated by stating that at the heart of healthy skepticism is being genuinely invested in seeking the truth regardless of how uncomfortable or undesirable the truth might be and having an aversion to falsehood (Le Morvan, 2011). And this fits perfectly with the spirit of science.

An important distinction between a healthy skeptic and the other kind, was made by the philosopher, Paul Kurtz, and later cited by Barry Beyerstein (Beyerstein, 1995, p.40) , in reference to what they called a methodological skeptic (which is what I call here a healthy skeptic). It is "one who, in approaching disputes, weighs the evidence fairly and accepts, provisionally, whichever position is best supported by logic and evidence." In other words, it is someone who provisionally (at least temporarily) accepts what is best supported by logic and evidence, rather than rejects the possibility that any truth can ever be found (which would describe a cynic) or believes anything right away (which would describe someone who is gullible). So if cynicism and gullibility are harmful to science, open-mindedness and critical approach to evidence are two skills, that we should continue to develop.

Part of being a healthy skeptic is being able to question and assess the validity of a claim. In chapter 3 we already discussed the difference between facts, claims of facts and opinions; the truth of the matter is that most often we are exposed to opinions or claims of fact; even scientific papers offer claims of facts since no single study is capable of answering questions without making errors or

leaving some lingering questions. What we need to do is to make judgements about presented evidence based on the quantity of the existing evidence and the quality of each single study.

Let's go through some of the things we can do to fact check a claim.

11.2 "Show Me the Data": Quality and Quantity of Scientific Evidence

Does the claim come supported with scientific evidence? If it does, you as a healthy skeptic, have to be open-minded and evaluate the quality and the quantity of the presented evidence. Ideally, you want to read the original source rather than trust the information that is relayed by another paper or a social media.

Remember the Strengths and Weaknesses of Each Research Method.

What we can draw from any single study depends on the nature of the research design. In other words, assuming that the ultimate goal of any scientific investigation is to offer an explanation or to find a cause and effect relationship, where it exists, an experimental study will have the highest internal validity (i.e, ability to establish cause and effect relationship). It draws its power from having control over all aspects of an investigation, especially its independent variable(s). That is, by directly manipulating the independent variable and controlling all extraneous variables constant, researchers can observe the direct effect of their manipulation. Thus, the cause and the effect, if it indeed is present, will be ascertained.

On the other hand, with the increase of control over conditions of a study, the degree to which the findings can be generalized to real life situations and people decreases. This happens because, as the control increases, conditions of a laboratory experimentation become more artificial and resemble less the reality of the phenomenon that is being investigated (i.e., external validity decreases). Take for instance, experimentation concerning drug intake using mice as subjects in research on drug addiction. Many such studies use rodents rather than human subjects in order to be able to manipulate the independent variable—e.g., drug intake (something they would not be able to do with human subjects for ethical reasons). By manipulating the independent variable, researchers increase their ability to detect the effect of the independent variable and increase the internal validity. But, at the same time, it decreases the external validity: humans and their environment are a lot more complicated. And so the artificiality of the experimental setting and, limitations of using non-human subjects, inadvertently decrease the extent to which researchers can generalize results of their experiments to people.

The opposite is true for non-experimental methods. For instance, research that surveys 50,000 people about their drinking habits may gain valuable insights into real people's attitudes, quantities of drinking, and maybe even events that precede the drinking (depending on the kinds of questions the survey includes); still a survey will not be able to establish the causal connection, even if certain significant correlations between reported preceded events or individual characteristics of the respondents and drinking do appear. So the findings must be treated as only correlational—that the relationship maybe causal or maybe not. The quasi-experimental research is typically treated as having a little bit of the benefits and the weaknesses of experimental and non-experimental studies.

How can these strengths and weaknesses be treated by a healthy skeptic when faced with a claim? Consider the claim and the source that is used to support it. If the claim is of a descriptive nature (e.g., how often people drink), then a published survey can be a great source of evidence granted it surveyed a good number of people. However, if the claim is made about the causes of people's drinking, no correlational research, including surveys, will be sufficient. Causal claims are the strongest if they are supported with studies using experimental methods.

Random Sample Selection.

External validity of any scientific investigation (experimental or non-experimental) depends on a careful selection of a representative and sufficient sample size. Both will influence the internal and external validity of the study; and thus, as a critical thinker and a healthy skeptic, you should always pay close attention to the description of the sample selection procedures typically presented in a methodological section of an article. The first concept refers to the principle that the sample must represent the population of interest (the people that the results are intended to be generalized to); while the second concept, the size of the sample, speaks to the importance of having a sufficient number of subjects in a study.

There are multiple strategies that a researcher can use from to select an unbiased sample. A commonly cited example that illustrates the importance of having a representative sample is the story of the Literary Digest magazine that failed to predict correctly the 1936 presidential election. The reason why this was a memorable failure was because up to that point the magazine had a strong reputation and was known for predicting every other past presidential election since 1920. Many articles and opinions had been written since then, trying to understand it. The majority of the critics believed that the error was caused by having a biased sample. Specifically, the Digest poll had sampled as many as 10 million people—an impressive sample size-- asking people who they would vote for in the mailed postcards (e.g., Crossley, 1937; Lusinchi, 2012). So the size of the sample was more than enough! However, the names of the respondents came primarily from the automobile registrations and telephone listings, which means that the majority of the

respondents in the sample were wealthy people and were more likely to vote for a Republican president. Therefore, the results were skewed because the sample didn't represent all likely voters. On the other hand, its new competitor, the Gallup group (founded in 1935 by Gallup, the founder of the American Institute of Public Opinion (AIPO)) sampled only a few thousand people but the strategy, as was written by one of the fellow pollsters, was based on a scientific method to "set up an America in microcosm" (Crossley, 1937), or in other words, to have a representation of the entire country. In sum, the success of the Gallop was attributed to having had a more representative sample of the likely voters. A more recent reevaluation of the Literary Digest fiasco has also suggested that the Literary Digest must have had a biased response rate, which is that only a certain portion of the sample actually mailed their responses.

There are various factors or consideration that typically go into determination of a sample size. Since hypothesis testing is based on statistics, factors that go into calculation of the sample size are also statistical in nature. Specifically, they are based on a known population size and standard deviation (i.e., the size of the population of generalization and variability of the measured phenomenon), confidence level (how confident the researcher chooses to be in the results representing the true matter of the phenomenon; in statistics it is known as the level of significance and setting an acceptable p value), power to detect true difference/effect, and expected effect size, if known (in other words, , what is the expected effect size of the treatment or manipulation of the independent variable, based on previous research findings) (Kirby, Gebski, & Keech, 2002).

Furthermore, any given research design has its own nuances that should also go into the calculation of the sample size. For example, an experimental study will have a tighter control over the conditions of the study, and thus, can get away with a smaller sample size as compared to a non-experimental research, such as a survey. Finally, researchers must consider and weigh in potential ethical concerns, such as "do no harm while maximizing the benefits". For example, a clinical study that may test an experimental treatment should recruit just enough subjects to be able to generate reliable data but with the minimum number of subjects who will bear the burden of participation in a potentially dangerous research.

Statistical terms aside, the reason why sample size matters is because it will most certainly affect the power of the study to find the true effect (Marszalek, Barber, Kohlhart, & Holmes, 2011). Therefore, given the importance and the complexity of determining a proper sample size, at the very least, an investigator should explain the decision making process that went into determining the sample size based on the aforementioned factors.

Converging Evidence vs Evidence from Replication.

More is better in the case of assessing strength of an existing evidence. No single study is convincing enough in terms of addressing the question of interest, but if a study is part of a body

of numerous investigations, which, collectively, reveals the same effect, then the evidence can be said to be overwhelming (granted all evidence is of good quality as well). As Carl Sagan once stated, “extraordinary claims require extraordinary evidence”. Incidentally, this statement was a paraphrase of Laplace's principle, who once said that “the weight of evidence for an extraordinary claim must be proportioned to its strangeness” (Gillispie, Gratton-Guinness, & Fox R. 1999).

We can think of two types of bodies of evidence: replication and converging evidence. While both are crucial in advancing science, they serve slightly different (albeit related) purposes. Let’s start with replications. In the most simple and minimal sense, replication is to repeat a study. The purpose of replication is to make sure that the obtained results were not accidental but rather a correct outcome of the investigation, under the conditions of the study and with the given sample (Goodman, Fanelli, & Ioannidis, 2016). In other words, it’s an attempt to fact check a study. Replication can be done by the original author(s) or by someone else. Theoretically speaking, success of this replication depends on whether the results of the replication were sufficiently identical to the original study. However, practically speaking, it also depends on how closely the procedures of the original study were followed during the replication (Nosek & Errington, 2020). Specifically, replication depends on similar research setting, identical equipment, apparatus, procedures, surveys, instruments, treatments, and method of recording observations. This, in turn, will also depend on whether the original study had provided accurately all the necessary details of the methodology (i.e., from sample selection to operationalization of the constructs, etc.) in order to replicate the study by someone else. However, as straightforward as the idea of replication may seem to be, it is not that simple. Here are at least two commonly debated issues: the meaning and causes of the received outcomes. First, the meaning of the results can be evaluated based on whether they are statistically significant and/or whether the effect of the results is the same. For example, suppose a replication finds a statistically significant sex differences in empathy (that female participants scored higher on emotional empathy than their male counterparts) replicating another study. However, suppose the differences are not that great, unlike the effect found in the original study. Can we conclude that the replication was successful? I am afraid we don’t have a good answer just yet. Second, the question can be raised as to the causes of these differences in the found effect. Original authors may (and have) argue(d) that the procedures or the sample (e.g., size or diversity) in replications are not correctly followed which resulted in the disparities in the findings between the original and the replication. Of course, it is also possible that replications fail for more sinister reasons, for example, because the original study was flawed or worse yet, data manipulation (e.g., forging data). Others have questioned the expertise of the teams doing replications and quality of the replication studies (Gilbert et al, 2015).

In a study published in Science, a top-tier English language research journal with less than 2% acceptance rate, a published paper detailed an effort of hundreds of independent research groups who endeavor to replicate 100 psychology experiments that were sampled randomly from publications in 2008 from three top-tier psychology journals (OSF, 2015). The aim of the study was to conduct high-fidelity replications of the selected research studies using similar sample sizes

and procedures. Overall, fewer than half of the studies were found to produce statistically-significant effects, and very few studies replicated observations of the same magnitude or quality by tightly-controlled manipulation of the independent variable. A 2018 sequel study of 28 replications for social psychology studies of experiments was published in the top-tier journals Science and Nature from 2010-2015 (Camerer, et al, 2018). This set of replications found a larger proportion of statistically-significant findings than the prior replication study, but, again, controlled manipulation of the independent variable(s) produced weak effects on dependent variable observations as compared to the original studies. This was despite the fact that sample sizes were increased five-fold for each replication compared to the original. These failures to replicate the effects of the original studies or to replicate even significant findings is referred to as the "replication crisis". A multitude of publications have endeavored to explain what might have caused the replication crisis. Explanations include questionable research practices and perverse incentives, that inflate and bias publications, (Ioannidis, 2005 and 2008; Button, et al, 2013; Loken and Gelman, 2017) and occasionally, honest but significant mistakes (Herndon, Ash, and Pollin, 2013; National Academies of Sciences, Engineering, and Medicine, 2019) or even outright fraud (Fanelli, 2009). The good news is that these issues are debated and conscientious researchers agree that replications are necessary; most researchers are willing to and invite honest criticism and transparency. For example, it is becoming a common practice to submit data along with the manuscript for professional scrutiny.

Now let's turn our attention to obtaining converging evidence; it is referred to obtaining similar findings by conducting a study using a different method or measures/operational definitions. For example, if a survey of over 600 middle school children found a positive association between exposure to video game violence and aggression, and if an experiment of 70 subjects found that more aggression was detected in a violent condition, then a conclusion can be made that a converging evidence from two different studies point to a similar effect, that violent video games can increase aggressive behavior. The strength of comparing results of methodologically different studies is that, as we already discussed, each method has its strengths and weaknesses. An experiment can offer a more conclusive answer to whether a direct effect exists between an independent and dependent variable, i.e., between playing violent video games and being more aggressive; while a correlational study can collect data on more subjects and assess violence or exposure to violence in more realistic ways, such as by directly asking people what and how often they play certain video games. To quote, Anderson and Anderson (1996), "Different methods are likely to involve different assumptions. When a conceptual hypothesis survives many potential falsifications based on different sets of assumptions, we have a robust effect." (p. 742; quoted in).

To further unpack the aforementioned quote, let's use the construct of aggression and the claim that violent video games cause aggression. Aggression can be operationally defined and measured in a number of ways. For example, a survey by Shao and Wang (2019) operationally defined it in terms of four separate dimensions: physical (e.g., striking another person), verbal (e.g., getting into arguments), as pure anger (e.g., being a hothead), and as hostility (e.g., being overly suspicious of

strangers). The survey directly asked the participants to rank statements describing each of the dimensions on a 5-point scale of how likely it describes them. The scores on the statements, if ranked truthfully and accurately by the participants, represent a fuller picture of human real-life aggression. This particular study did find that exposure to violent video game content was positively related to the overall score on aggression (all above dimensions combined).

Now let's look at an experimental study by Hasan, Begue, Scharkow, and Bushman (2013), where aggression was measured two ways. First, it was measured by assessing participants' hostile expectation bias, which is a well-known psychological concept when one expects that others react to conflict with aggression. People who hold this bias are also more likely to react aggressively as they view others as more hostile. The second way was to have participants complete a computer game where they had to administer a noise blast through a pair of head-phones to their losing partner (the partner was a confederate). Aggression in this game was determined by the intensity and duration of the noise.

The hostile expectation bias was assessed by having the participants completed a story where a driver crashes in to the main character's car causing a significant damage. Participants were asked to decide what happens next in terms of what the main character would do or say in response. Comparing to the survey, aggressive behavior was certainly measured in more narrow ways. The result of the study, by the way, did find that the hostile expectation and aggression increased only in subjects in the experimental condition. While by itself, the survey by Shao and Wang (2019) and the study by Hasan et al. (2013) provide a more or less convincing evidence of a negative effect of violent video games, together, they add a lot more converging evidence, compensating each other for methodical limitations and by complementing the assessment of aggression using different measuring tools.

11.3 Science vs Pseudoscience

A healthy skeptic should be able to distinguish science from something that is non-scientific but masquerades or calls itself as a scientific enterprise—we call it pseudoscience. Pseudoscience may use scientific research only selectively or simply use scientific terms to create an appearance of legitimate science. Some of the common examples of pseudoscience are astrology, parapsychology, UFO abduction studies, etc. Although, intuitively, many people can recognize pseudoscience for its outlandish or dubious claims, still many people continue to believe in pseudoscientific claims, and some fields may fall in so called gray area of something between science and pseudoscience. So, to help all critical thinkers and healthy skeptics to discern junk science from true science, it is useful to have a more specific set of descriptive characteristics to distinguish science from pseudoscience. Mario Bunge, a laureate of the International Academy of Humanism and a Philosophy, came up with the following system of features that can be of use in

recognizing pseudoscience. He first categorizes science and pseudoscience into two groups: research field and belief field, where the research field stands essentially for fields that use some sort of intellectual or scientific work to make inquiries, while the latter is based on a set of beliefs only. Religion, political ideologies, pseudoscience and pseudo-technologies can thus all be placed into the belief field. To elaborate, he lists the features that describes the research or science field, and he argues that all must be met to be considered a science or a research field.

Here are the features that characterize a research field according to Mario Bunge that we can all use (Bunge, 1991):

A research field will have:

1. A Research community (C)—a social system of people with professional training in the area of their research connected through work, communication, exchange of ideas, etc.
2. A Society (S)– hosts and encourages the above research community with certain culture, organization and economy
3. A Domain (D) – domain or universe of discourse of ideas rather than freely floating ideas
4. A General outlook or philosophical background (G) -- (a) grounded in ontology of realistic and lawful accounts of the world with its changes (or no changes) according to some sort of laws, (b) a realistic theory of knowledge, (c) "a value system", (d) "the ethos of the free search for truth"
5. A Formal background (F)—mathematical or logical tools that are used in inquiries
6. A Specific background (B)—assumptions or presuppositions (related to D) borrowed from other fields
7. Problematics or set of problems (P)—research questions or problems that the field can address
8. A Specific body of accumulated knowledge (K)—body of knowledge that is accumulated as the result of scientific explorations
9. Aims or goals (A)—discovering and using laws related to (D)
10. Scientific methods—methods of inquiry are scientific (e.g., testable)

So, if in doubt about a particular field, check it against the 10 aforementioned features to see if it meets them. If it fails some of them, it may fall into that gray area, and if it is missing most or all of them, it is a full-blown junk science and you should stay away from it!

One key difference between science and pseudoscience is the community of people working collaboratively and transparently, fact-checking each other through peer-review processes. This helps those field to progress. That is not to say that at times, scientists can be stubborn and even resistant to new theories. Perhaps the best example of this is the idea that the earth revolves around

the sun. However, eventually this so called crazy idea was proven to be true. So, it is up to the individual proposing a novel idea to present the evidence and to convince the scientific community of its correctness (Beyerstein, p. 6). Those visionaries "brought the initially doubting fields around by force of evidence, not mere conjecture and special pleading" (p. 6).

A belief field doesn't change or changes little and/or as a result of controversy, some sort of force or revelation. Unlike scientists, pseudoscientists tend to work alone. They operate outside the mainstream, and do not collaborate or communicate with experts in the field. They do not build connections and relationships, most likely for the fear of being debunked. Their claims are non-falsifiable. They often misuse or misrepresent data to fit their narrative. Pseudoscientists are not eager to seek out new evidence, and do not self-correct, because for them, their beliefs tend to be personal. While on the other hand, pseudoscience asserts itself as a legitimate field, on the other hand, it doesn't want to follow the rules that science abides by (Beyerstein, 1995). To give an example, claims have to be falsifiable and tested, and evidence must be replicated. Finally, oftentimes pseudoscientific claims offer comforting views which probably explains why many people continue to buy into pseudoscientific claims. For example, try psychic reading. Chances are, the reading is going to be very uplifting and positive, whereas science can be rough (Beyerstein, 1995).

Another reason why people gravitate towards pseudoscientific claims is because they are easy to understand and they tend to promise a whole lot, e.g.,, holding crystals will cure your depression, hydroxychloroquine will cure you of Covid-19. Science, on the other hand, produces fuzzy results, full of nuances and caveats; the so called facts are almost never settled. Moreover, no single study will ever satisfy you with a direct answer, as we discussed earlier. This must be frustrating to some people.

11.4 Conspiracy Theories

What is a conspiracy theory and what does it have anything to do with science and critical thinking? Let's start with a definition. It can be defined as "a subset of false beliefs in which the ultimate cause of an event is believed to be due to a plot by multiple actors working together with a clear goal in mind, often unlawfully and in secret" (Swami & Furnham, 2014, p. 220). A recent survey found that half of the Americans believed in at least one conspiracy theory (Oliver & Wood, 2014). This chapter was written during the coronavirus pandemic, in the summer of 2020. Dealing with this pandemic was particularly difficult because some people refused to believe that we even had a pandemic; for example, one conspiracy theory was that Democrats over-exaggerated the threat of the corona virus to hurt Trump, the Republican president, politically. There were also various extravagant theories about the origin of the virus. For instance, one was that Bill Gates had created the virus and the vaccine, in order to make lots of money as soon as the virus would spread

over the world and people would become desperate for cure or a vaccine (By the way, this reminds me of the plot from the Mission Impossible 2 movie. Incidentally, many if not all Hollywood thrillers resemble a conspiracy theory by setting up the "bad guy" to be one or a group of powerful people working to conspire in order to make a bunch of money and/or destroy the rest of the people.). Another theory was that some very important people at the top [of course we don't know who they are] had decided to reduce the population [because the earth can't sustain so many people] by developing the corona virus, to affectively kill millions of people [just like the Spanish flu in 1918.].

So, what does this all have to do with critical thinking? And, perhaps even more importantly, why do people believe in conspiracy theories? Thankfully, there have been some research conducted on this topic which can help shed some light. In a review paper by Douglas, Sutton, and Cichoka (2017), researchers summarized most current understanding about what attracts some people to conspiracy theories. Specifically, there seems to be three motivational, evolutionary rooted, drives: epistemic (desire for knowledge and causal explanations of things), existential (innate drive for control and security), and social (to make oneself and one's group look better and/or to attribute bad outcomes to others). Let's unpack each one of them.

An innate human curiosity and the need for causal understanding of things, which we had discussed in chapter 1, drives some people to conspiracy theories, particularly in the absence of solid answers. Unlike scientifically derived answers, which come with nuances and degrees of uncertainty, conspiracy theories give us straightforward answers (just like pseudoscientific claims) (By the way, since conspiracy theories are easy and more fun "to swallow", this can explain why Hollywood blockbuster films are extravagant conspiratorial stories). Having answers also gives us comfort which satisfies our second drive for control and security. For instance, as pandemic produced fear in people, conspiracy theories gave some people control and security.

Here is a real life example that I came across in an article in Vox (https://www.vox.com/first-person/2020/5/15/21258855/coronavirus-covid-19-conspiracy-theories-cancer?fbclid=IwAR1VmE3ZraPAG9ugn056wQEM5OHZLwHpgaO3CYg_W4DBfs01EdSZmkOPJKM). It was written by a woman, Dannagal G. Young, who lost her husband to cancer. In the article she describes her journey from grieving, after discovering about her husband's tumor, to becoming obsessed with searching for answers and trying to figure out what caused the cancer. As she writes, she became obsessed with various conspiracy theories after her husband's doctor suggested that her husband's cancer was a random occurence. The absence of an actual cause to blame her husband's illness for, fueled her search for the answers, and she found them in conspiracy theories. She writes: "These words [randomness] were not comforting to me.

Dannagal writes further this:

> *"Each time I landed on a possible culprit, my anger reenergized me. Instead of making me feel hopeless, it gave me a target and suggested there might be some*

action I could take. If it were from his work or from an old factory site, maybe I could file a lawsuit. Maybe I could launch an investigation or trigger some media exposé. If I could just find the right person or thing to blame, I could get some justice. Or vengeance. Or … maybe just a sense of control."

This mixture of emotions—anger, sadness, sense of control and justice--illustrates two motivational drives, existential and social—the drives to gain some control over an uncontrollable situation and to assign blame/punishment to someone or something for the pain it caused it her and her husband.

Research indeed suggests that people are particularly vulnerable to be drawn to conspiracy theories when (1) there is a lack of clear explanation, (2) when under stress (e.g., due to uncertainty), and when events are "large in scale and mundane explanations just don't seem satisfactory enough (Douglas, Sutton, & Cichoka (2017). In addition, there are individual characteristics that make some people more susceptible to conspiracy theories than others, such as being less tolerable of randomness and uncertainty of life, and having a strong need for cognitive closure. Both traits are related to the aforementioned epistemic drive to seek answers. However, it appears that people also vary in the degree to which they can tolerate the degree of uncertainty and not knowing things "for sure" (Marchlewska, M., Cichocka, A., & Kossowska, M. (2017). Scientists, for instance, tend to be more tolerable of uncertainties since they must constantly deal with uncertainties and revisions when it comes to new findings and facts (i.e., by default, science is never complete and conclusive). People on the other end of the spectrum of cognitive closure are more likely "to leap" to conclusions or judgements because they require immediate closure (Kruglanski & Webster, 1996).

Going back to the belief in conspiracy theories, they clearly provide people with immediate gratification for cognitive closure and help them have someone or something to hold responsible for the negative events (which is, typically, a more powerful group) (Kofta & Sedek, 2005). Of course, it is also conceivable that some conspiracy theories maybe true. As I already mentioned, we should remain open-minded at all times. Then, how can we recognize a conspiracy theory from a plausible one? Surprisingly, conspiracy theories are so diverse that even scholars and social scientist haven't come up with one set of definitive characteristics that would describe all conspiracy theories and help us distinguish a plausible one from a conspiratorial. However, a few principles can be applied. First, consider the principle of parsimony. This principle is typically associated and credited to William of Ockham, a fourteenth century English philosopher. Parsimony refers to the idea that simplicity of an explanation should be favored above any more complicated ones. As one psychologist and philosopher, Robert Epstein elaborated: " Where we have no reason to do otherwise and where two theories account for the same facts, we should prefer the one which is briefer, which makes assumptions with which we can easily dispense, which refers to observables, and which has the greatest possible generality." (Epstein, 1984). Second, you can examine the fact pattern in a conspiracy theory in a logical or scientific manner, for example, recognizing illogical arguments, unproven facts, and the nature of connectedness of

events (e.g., whether the events are connected at all or connected by chance). Most conspiracy theories are built on a chain of complicated, thinly connected events. These events maybe spuriously correlated or may not even be connected at all. Third, you should note whether a theory fails the principle of falsifiability (Banas & Miller, 2013; Swami, Voracek, Stieger, Tran, & Furnham, 2014). A non-falsifiable conspiracy theory is the one where even if you try to present counter-arguments or facts, a conspirator will easily dismiss them as "fake news" or "misinformation", in other words, by suggesting that you are being "brainwashed" or misinformed by the same powerful group of people who are at fault. In other words, no amount of evidence or persuasion will ever prove the theory to be false. A critical thinker should recognize that as a major red flag.

Finally, now that you know what credible scientific evidence is, consider how any given theory relates to and deals with new evidence. A successful theory should be able to generate successful predictions (New evidence should most certainly emerge eventually). A conspiracy theory fails to generate successful predictions; furthermore, often time even initial conditions in a conspiracy theory must be modified to accommodate emerging contradictory evidence (e.g., Clarke, 2002).

Concluding Remarks

If there is one main takeaway from the research methods course and this book is that science is a difficult and thorny process. It is also gradual, with more failures than successes. The breakthroughs don't normally come out of one single study, but out of an accumulation of many investigations, and over time. I hope you have acquired a better appreciation for the scientific process and the people who devote their entire professional lives to advancing human knowledge and making our every-day lives more enjoyable. Let's appreciate expertise. As for your personal development, even if you are not planning on becoming a scientist yourself, you can benefit by following scientific principles and critical thinking to have a more accurate understanding of the world.

References

Banas, J. A., & Miller, G. (2013). Inducing resistance to conspiracy theory propaganda: Testing inoculation and metainoculation strategies. Human Communication Research, 39, 184-207.

Beyerstein, B. L. (1995). Distinguishing science from pseudoscience. Prepared for the Center for Curriculum and Professional Development. Department of Psychology, Simon Fraser University.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14, 365

Bunge, M. (1991). What is science? Does it matter to distinguish it from pseudoscience? A reply to my commentators. New Ideas in Psychology, 9, 245-283.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour, 2(9), 637–644.

Clarke, S. (2002). Conspiracy theories and conspiracy theorizing. Philosophy of the Social Sciences, Vol. 32 No. 2, June 2002 131-150

Crossley, A. M. (1937). Straw polls in 1936. Public Opinion Quarterly 1, 24-35.

Epstein, R. (1984). The principle of parsimony and some applications in psychology. The Journal of Mind and Behavior, 5, 119-130.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PloS One, 4(5), e5738.

Gillispie C. C., Gratton-Guinness I., Fox R. (1999). Pierre Simon Laplace, A Life in Exact Science. Princeton, NJ: Princeton University Press

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? Science Translational Medicine, 8(341), 341ps12.

Young, D. G. (2020, May 15). I was a conspiracy theorist, too. Vox. https://www.vox.com/first-person/2020/5/15/21258855/coronavirus-covid-19-conspiracy-theories-cancer?fbclid=IwAR1VmE3ZraPAG9ugn056wQEM5OHZLwHpgaO3CYg_W4DBfs01EdSZmkOPJKM

Hasan, Y., Begue, L., Scharkow, M., & Bushman, B. J. (2013). The more you play, the more aggressive you become: A long-term experimental study of cumulative violent video game effects on hostile expectations and aggressive behavior. Journal of Experimental Social Psychology, 49, 224-227.

Herndon, T., Ash, M., & Pollin, R. (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. Cambridge Journal of Economics, 38(2), 257-279.

Ioannidis, J. P. (2005). Why most published research findings are false. PLoS Medicine, 2, e124.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. Epidemiology, 640-648.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. Science, 355, 584–585.

Le Morvan, P. (2011). Healthy skepticism and practical wisdom. Logos & Episteme II, 1, 87-102

Lusinchi, D. (2012). "President" Landon and the 1936 "Literary Digest" Poll: Were automobile and telephone owners to blame? Social Science History, 36, 23-54.

Kirby, A., Gebski, V., & Keech, A. C. (2002). Determining the sample size in a clinical trial. Medical journal of Australia, 177(5), 256-257.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "Seizing"and"freezing". Psychological Review, 103, 263–283. http://doi.org/10.1037/0033-295X.103.2.263

Machuca, D. E. (2011). Ancient skepticism: Overview. Philosophy Compass 6, 4.

Marchlewska, M., Cichocka, A., & Kossowska, M. (2017). Addicted to answers: Need for cognitive closure and the endorsement of conspiracy beliefs. European Journal of Social Psychology. Advance online publication. doi:10.1002/ejsp.2308

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in Psychological research over the past 30 years. Perceptual and Motor Skills, 112, 331-348.

National Academies of Sciences, Engineering, and Medicine. (2019). Reproducibility and replicability in science. National Academies Press.

Nosek, B. A., & Errington, T. M. (2020). What is replication? PLOS Biology, 18(3), e3000691.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic vs. analytic cognition. Psychological Review, 108, 291–310. http://dx.doi.org/10.1037//0033-295X.108.2.291.

Oliver, J. E, & Wood, T. J. (2014). Conspiracy theories and the paranoid style(s) of mass opinion. American Journal of Political Science, 58, 952-966.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716–aac4716.

Shao, R., & Wang, Y. (2019). The Relation of violent video games to adolescent aggression: An examination of moderated mediation effect. Frontiers in Psychology, 10.

Squire, P. (1988). Why the 1936 Literary Digest poll failed. The Public Opinion Quarterly, 52, 1, 125-133.

Swami, V. & Furnham, A. (2014). Political paranoia and conspiracy theories. J.-P. Prooijen, P.A.M. van Lange (Eds.), Power politics, and paranoia: Why people are suspicious of their leaders, Cambridge University Press, Cambridge (2014), pp. 218-236

Swami, V., Voracek, M., Stieger, S., Tran, U. S. & Furnham, A. (2014). Analytic thinking reduces belief in conspiracy theories. Cognition, 133, 572-585.



SENTIA PUBLISHING

ISBN 978-1-7359340-2-0

9 781735 934020 90000>