Title

Exploring the Decision Making Process in Statistical Data Analysis: A Qualitative Study of Quantitative Researchers

Author

Ho, Timothy

Publication Date

2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Exploring the Decision Making Process in Statistical Data Analysis:

A Qualitative Study of Quantitative Researchers

A dissertation submitted in partial satisfaction

of the requirements for the degree Doctor of Philosophy

in Education

by

Timothy Ho

2015

ABSTRACT OF THE DISSERTATION

Exploring the Decision Making Process in Statistical Data Analysis:

A Qualitative Study of Quantitative Researchers

by

Timothy Ho

Doctor of Philosophy in Education

University of California, Los Angeles, 2015

Professor Christina A. Christie, Chair

Quantitative data analysis is a cognitively demanding process. Inferences from quantitative analyses are often used to inform matters of public policy and to learn about social phenomena. However, as statistical analysis is typically conducted behind closed office doors, little is known about how analysts decide on the final statistical model that important policy decisions rely upon for determining the effectiveness of programs and policies. As social programming becomes increasingly reliant on quantitative data analysis, it becomes imperative to examine the quality of information stemming from these sources of evidence. This project presents the results of a qualitative research project that explores the cognitive processes of quantitative data analysts.

Seven quantitative analysts participated in the study. Participants were interviewed and observed during the course of analyzing a quantitative data set to uncover underlying cognitive processes used during data analysis. A typology of the decisions encountered by quantitative analysts in the sample is presented. A framework for decision making in quantitative analyses is presented that is borrowed from the field of cognitive psychology. Metacognition in quantitative data analyses is also explored.

Findings suggest that despite prescriptive procedures that are intended to facilitate objective analysis, quantitative analysts are frequently subject o external influences that may affect the results of quantitative analyses. Social factors and context influence how quantitative researchers make decisions during data analysis which call into question the objective nature of quantitative analyses. Quantitative analysts use story telling as a method of making sense of the findings from quantitative analyses. Implications for practice and for the teaching of quantitative methods are discussed.

The dissertation of Lisa Marie Dillman is approved.

Marvin C. Alkin

Michael Seltzer

Todd Franke

Christina A. Christie, Committee Chair

University of California, Los Angeles

2015

For my mother and father, who have sacrificed their lives so that I could fulfill mine;

my wife, Elizabeth, for this could not be accomplished without her patience

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

This project would not have been completed without the support and backing of many people that have helped me over the years. First, I cannot properly express enough gratitude to my advisor, Tina Christie, who first believed in me and gave me the chance. Her guidance and patience was just what I needed to succeed and I could not imagine being to make it without her. I want to give thanks to the rest of my committee: Marv Alkin, Mike Seltzer, and Todd Franke. I could not have asked for a better committee to help steer my dissertation. Their thoughtful comments certainly helped to make this project realize its full potential.

Other professors at UCLA Social Research Methodology have no doubt positively impacted me. In no particular order, Noreen Webb, Jose-Felipe Martinez, Mike Rose, and Katie Anderson Leavitt have all contributed to the development of this project and to my development of a scholar. Thanks to GSEIS student support staff, particularly Amy Gershon for helping me navigate the web. I will miss running into everyone in Moore Hall.

Thank you to the participants of my study, those that have taken the time out of their busy schedules to help with my project. I asked a lot from you, and I appreciate that you showed interest in my project and took the time to rearrange schedules to accommodate me.

I want to express my sincerest appreciation to the other Social Research Methodology students that have provided me with the positivity that I needed to accomplish these goals. Your presence has given me the proper mental health to stay focused. In particular, Lisa showed the way out of the woods. Alejandra was the greatest TA partner and an outlet for my passive aggressiveness. Mark was an all around smart guy that helped me focus this project loaned me his cap and gown. Belinda and Rebecca participated in the writing group that started as the first step. Patty, Debbie, & Celina helped make evaluation fun and stimulating. Jenn and Minh gave the Ho's strength in numbers. There are many others have I have failed to mention, but I am forever grateful for all the time we have spent together, laughing and commiserating over drinks. Graduate school is supposed to be a stressful and demanding life event, and indeed it was, but I would do it all over again if I could have you all by my side once more.

Thanks to everyone that I had the honor of serving as a teaching assistant for during the 230 series. You were an inspiration to this project and I hope that I helped you with yours as well.

To the great state of Oregon that has given me the mental clarity that I needed to complete this project. I could not ask for a more beautiful and inspiring place to write.

Finally, I want to once again thank my family. In particular my mother and father who have instilled me with the value of hard work and given me everything I needed to succeed. They have given so much to provide me with a better life, and I have worked hard in life just so I can make them proud. Adela and Mike have also given me the support over the years and have believed in me. Finally, I need to thank my wife, Elizabeth, who has stood by me as I entered this journey. We have gone through a lot over the years and I could not have asked for a better companion to accompany me through graduate school and through life.

It has been my privilege to have known and worked with all of you. I gave it my all, and I hope that I have not let anyone down.

VITA

EDUCATION

2009 M.A., Education
University of California, Los Angeles, CA

2004 B.A., Psychology
California State Polytechnic University, Pomona, CA

PROFESSIONAL EXPERIENCE

2014-2015 Research Analyst
Department of Human Services
State of Oregon, Salem OR

2011-2014 Teaching Assistant
Graduate School of Education and Information Studies
University of California, Los Angeles

2009-2014 Graduate Student Researcher
Graduate School of Education and Information Studies
University of California, Los Angeles

2006-2009 Research Associate
EMT Associates
Encino, CA

SELECTED PUBLICATIONS

2013 Luskin, R. J. C., & Ho, T. (2013). Comparing the intended consequences of three theories of evaluation. Evaluation and Program Planning, 38, 61-66.

2012 Franke, T. M., Ho, T., & Christie, C. A. (2012). The Chi Square test: Often used and more often misinterpreted. American Journal of Evaluation, 33, 448-458.

SELECTED PRESENTATIONS

2013 Ho, T. (October 2013). Strategies for Improving the Quality of Quantitative Data. Paper Presented at the annual meeting of the American Evaluation Association. October 16, Washington, DC.

2011 Ho, T. (November 2011). A review of fidelity of implementation measures in the What Works Clearinghouse. Paper Presented at the annual meeting of the American Evaluation Association. November 4, Anaheim, CA.

2010 Franke, T., Ho, T., & Christie, C. (November 2010). The Chi-square test: often used and more often misinterpreted. Paper Presented at the meeting of the American Evaluation Association. November 11, San Antonio, TX.

2010 Ho, T. (November 2010). Using logic models to facilitate comparisons of evaluation theory. Copresented with Tanner Wallace, Mark Hansen, Anne Vo, Lisa Dillman, and Rebecca Cohen. American Evaluation Association. November 11, San Antonio, TX.

2008 Ho, T. (March 2008). Effective and culturally sensitive prevention and early intervention strategies working with immigrant Asian youth. Co-presented with Rocco Cheng, Liyu Sui, and Simon Wai. Cultural Competence and Mental Health Summit XV. March 25-26, San Diego, CA.

2007 Petrucci, C., & Ho, T. (October 2007). A study of DUI offenders: preliminary results of an alcohol problem computerized screening. Presented at the Association of Criminal Justice ResearchersCalifornia. October 11th, 2007. Long Beach, CA.

2005 Armenta, B. E., Willis Esqueda, C. E., Ho, T., & Bivens, T. (2005). Stereotype threat and stereotype lift: The moderating role of group identity. Poster session presented at the 17th annual meeting of the American Psychological Society, Los Angeles, CA.

SELECTED HONORS AND AWARDS

2013-2014 Dissertation Year Fellowship, UCLA Graduate Division

2012-2013 Distinguished Teaching Assistant/Special Reader Award, Graduate School of Education and Information Studies, University of California, Los Angeles

CHAPTER 1

INTRODUCTION

The goal of most social science research is to provide information that will benefit society and lead to improved policy making. One such source of information is the statistical analysis of quantitative data. Using various quantities of data, statisticians and quantitative researchers in various academic disciplines analyze data that have been collected to begin to understand the patterns that exist in the particular settings being examined (Bartholomew, 1995). Using the apparently objective information provided by such analyses, social science researchers begin to infer whether the basis of any variance can be attributed to sources that can be manipulated, either through policy decisions or other social programming, thus providing an agent to bring about positive change in the real world. Altogether, it is a process whereby analysis of an experimental or near experimental situation is prescribed by a formal methodology, and the results inform public policy for the sake of social betterment (AbouZahr et al., 2007; Henry, 2000; Scott, 2005).

Increasingly, information derived from statistical data analysis has become a basis for making social policy decisions. The movement toward evidence-based practices has expanded from its origins in health-based research and clinical trials into other spheres (Boruch, 2005). Recent efforts of the Cochrane Collaboration and Campbell Collaboration, repositories of evidence-based health and social science research, come to mind as examples. This trend has prioritized information from randomized controlled trials and quantitative data to answer questions concerning cause and effect. These types of data ostensibly provide objective information from which to form the basis of matters of public policy. As such, it becomes more important to consider issues of validity with these types of research, as social science and educational research increasingly moves towards using these types of data and information to inform policy decisions.

As educational research and other fields move towards more evidence-based practices—and using statistical data as the definition of "evidence"—it becomes more important to examine these methods at a more reflective level. The arguments for using statistical data are compelling; they provide a much more objective measure in determining the effectiveness of an intervention or the quality of a teacher than methods that rely on observation or testimonials—sources of data that can be easily subject to bias. However, if policy decisions more heavily lean on statistical results, an examination of the quality of evidence from these statistical models is needed. In short, the quality of decisions made is only as good as the quality of evidence used to support those decisions.

Statistical data analysis is undoubtedly a cognitive process requiring a certain level of expertise to perform such analyses. There are many decisions to be made in the course of a typical analysis and these decisions require the analyst to assess the situation, troubleshoot the problem, make judgments, and decide on the appropriate course of action that will best lead the analyst to the goal of a final model. There is also an evaluative component that governs this process, in assessing whether each decision made was indeed the best course of action to complete the analysis. This is often described as metacognition, or thinking about one's own thought processes (Veenman, Van Hout-Wolters, & Afflerbach, 2006). The cognitive and metacognitive processes used by quantitative researchers have seldom been explored. Like any cognitive process, quantitative data analysis is subject to variation in individual differences in cognition. Further, it stands to reason that the results from such analyses may also vary as a result. This has implications for the use of statistical results, particularly their ability to inform policy decisions. The purpose of social science research and evaluation is to offer information to policy makers to decide how to program society (Henry, 2000; Weiss, 1988, 1999). However, if this information is affected by variation in cognitive processes, then its ability to provide objective information to policy makers is hampered.

This dissertation research was designed to explore the human elements that exist in quantitative research that uses formal statistical theory and methodology. Using qualitative methods, the research explored the decisions made by quantitative researchers during data analysis and how those decisions are made. To borrow parlance used by quantitative researchers, the proposed research attempted to understand the extent to which variation exists in the decisions made during quantitative research and explain any variance that was discovered. Understanding these sources of variance helps to reveal the extent to which inferences made from statistical models, and in turn the decisions that are derived from such models, are free from the influences that go into decisions made by quantitative researchers.

More specifically, the proposed research will explore the decision making that occurs by quantitative analysts during the course of data analysis. Human cognition is a vital component of quantitative analysis as evidenced by the required human user sitting in front of the computer. Unlike, for example, filing a 1040 EZ tax form by imputing numbers from a W-2, a researcher does not simply provide a statistical software package with numbers and immediately get something of substance in return. It is not an automated process whereby the user simply feeds the computer software data tables and receives results as an end product. Rather, there are steps between inputting data and interpreting results, some that are prescribed by statistical methodology and others that are more subjective. Borrowing from methods used to study decision making and expertise, the current study used qualitative methods to study the cognitive processes that quantitative analysts engage in during the course of statistical analysis. In short, the present research intended to unravel the black box of the mind that exists during quantitative data analysis by exploring the following research questions:

(1) What types of decisions are made by quantitative researchers during the data analysis process?

(2) What are the cognitive and metacognitive processes of quantitative researchers during the course of data analysis?

(3) What types of information do quantitative analysts rely on to solve problems and make decisions during data analysis?

(4) To what extent do these cognitive and metacognitive processes differ across disciplines?

Next, Chapter 2 provides a summary of relevant research that served as a foundation for the current research project. It also introduces conceptual frameworks that are used as a basis to discuss the findings of the research in later chapters. Chapter 3 provides details on the data collection and analytical strategies used to explore the research questions. Then, the research findings are distributed across three chapters. Chapter 4 provides a description of the type of decision made by study participants and the factors that influenced those decisions. Chapter 5 discusses the metacognitive strategies used by researchers to evaluate the efficacy of the decisions during quantitative data analysis. Finally, Chapter 6 summarizes the research findings and discusses implications for current and future quantitative analysts, as well as provides suggestions for future research.

CHAPTER 2

REVIEW OF RELEVANT LITERATURE

A primary concern of any social scientist, whether in psychology, education, public health, or any other discipline, is to uncover relationships in the world. This is done by generating theory and developing hypotheses about the world and then collecting data in a controlled setting to see whether these hypotheses hold up in the sample under consideration. These findings are then potentially used to inform decisions of public policy that will hopefully move towards social betterment.

Although social scientists always strive to uncover meaningful relationships, and they typically put safeguards and precautions into place to ensure that false inferences are avoided, errors inevitably make it through the screening process. The sources of these errors can be difficult to detect, as they can occur at any step of the research process. Their presence, however, impacts the validity of the information generated from the research and the evaluation of social programs, and this in turn affects the quality of policy decisions being made. Thus, as policy making becomes increasingly reliant on evidence-based practices, it is important to understand the sources of cognition that may lead to potential errors.

In the educational context, the move towards evidence-based practices is apparent in the No Child Left Behind Act of 2001 (NCLB, 2002) and is manifested in the creation of the What Works Clearinghouse (WWC), sponsored by the U.S. Department of Education and the Institute of Education Sciences (Schoenfeld, 2006; Gersten & Hitchcock, 2009). The WWC is a repository of educational interventions that have been shown to be effective through statistical analysis. To be clear, a major criterion for inclusion into the WWC is a rigorous research design (i.e., randomized controlled trials) to test the effectiveness of the intervention under consideration. This implicitly favors the use of statistics and quantitative data to analyze the effectiveness of interventions.

Another pertinent example of the increasing use of quantitative data to inform decision making is the rise of teacher evaluation systems (Campbell, Kyriakides, Muijs, & Robinson, 2004; Little, Goe, & Bell, 2009). Teachers have recently come under more scrutiny regarding their effectiveness in the classroom, as many have argued that teacher quality is the strongest and most controllable element that explains variation in student achievement. As such, state and federal policy has moved towards models that use student standardized test scores to measure teacher effectiveness. Previous teacher evaluation models relied on other sources of information such as peer reviews or principal observations. In other words, the qualitative information—i.e., peer reviews and classroom observations—that has been traditionally used to evaluate teachers is evolving towards a model that more heavily relies on quantitative information.

Statistical analysis and quantitative data in educational settings are often described as being more objective and less susceptible to bias than other forms of data (Rockoff & Speroni, 2011; Watanabe, 2011). Procedurally, a statistics textbook can provide details regarding the sequential steps in such data analyses that anyone with technical expertise can perform. This provides a foundation for making unbiased inferences that will later inform matters of social policy and programming. However, not all decisions and steps required in statistical analyses are covered in any single textbook. This may lead to variations in how analysts perform statistical analyses, which, in turn, may potentially be sources of error.

In the following sections, a discussion is presented on the variations that can exist in social science research, particularly in designing evaluation and research studies. The implications of these variations are discussed with regards to their effects on the inferences being generated and for policy making. Then, previous theoretical work is presented that describes how variations can exist in the course of data analysis. Finally, constructs from cognitive psychology are used to help frame the current project.

Validity in the Research Process

Decisions made during the research design and analysis phases affect the validity of research findings. Shadish, Cook, & Campbell (2002) defined validity as the "approximate truth of an inference" (p. 34). The primary goal of any research or evaluation study is to provide evidence that is as valid as possible, and the decisions made during research and evaluation studies affect the validity of inferences drawn from such studies.

Shadish et al. (2002) distinguished four types of validity in their typology. *Internal validity* is concerned with the accuracy of "inferences about whether observed covariation between A and B reflects a causal relationship from A to B in the form in which the variables were manipulated or measured" (p. 53). This refers to whether the research or evaluation study is actually able to make claims about the relationships under study given the study design elements. *External validity* refers to whether the inferences made regarding relationships between variables in a given study can be generalized to other settings in the population. *Construct validity* involves whether the sampled units and measures in a given study represent the higher order units and constructs as intended. Finally, *statistical conclusion validity* refers to the appropriate use of quantitative methods in order to make inferences about the relationships between variables.

There has been much theoretical work regarding the threats to these types of validity and the research and evaluation design choices that affect validity. For example, the decision to employ randomized controlled trials may increase internal validity by yielding unbiased causal effect estimates for specific interventions (Shavelson & Towne, 2002; Bickman & Reich, 2009). In addition, seminal work by Cronbach and colleagues pointed to the many factors that may affect construct and external validity (Cronbach & Meehl, 1955; Cronbach, 1982). With regards to statistical conclusion validity, any textbook or journal article that involves statistical methods is inherently concerned with generating valid conclusions from quantitative analyses. Shadish et al. (2002) listed the many threats to statistical conclusion validity, including low statistical power, unreliable measures, inflation of error rates, and violation of assumptions of statistical tests.

Although they were mainly concerned with the design stage of research and evaluation, Shadish et al. (2002) also noted possible solutions to these threats to statistical conclusion validity. For example, possible remedies included increasing sample size or using matched or stratified samples to obtain sufficient statistical power; selecting measures that are more reliable and do not restrict the range of responses; or increasing the strength or variability of treatment groups to detect differences among these groups. Less attention was given to choices that occur during the course of data analysis, when researchers are ostensibly past the point when design decisions can be adjusted. I would argue that the decisions made in fitting statistical models are also a potential threat to statistical conclusion validity, although this was not explicitly addressed by Shadish and colleagues.

The implications of choices made during data analysis may be apparent in how the results of such analyses are used. More often than not, the findings of statistical analyses are used to inform decisions in the policy or social arena. By showing whether the intervention is affecting key outcomes, results from statistical tests can help policy makers decide whether to keep implementing an intervention or to scrap it altogether. At a local level, a school superintendent of a small school district may use quantitative data to decide whether to pursue e-books as an option for the classroom. Data showing that the use of e-books increased student achievement may persuade the superintendent that it is appropriate for the district. It is clear that quantitative information helps inform decision makers at all levels to make choices that are supported by statistical evidence.

Furthermore, the information provided by quantitative data is presumably more objective than other modes of inquiry and is thus perceived to be more credible. The quantitative data that go into statistical models are arguably less susceptible to bias than observational methods or testimonials. In this line of reasoning, the results from these statistical models should also contain less bias, thus allowing policy makers to make decisions that are equivalently free from bias. However, if there is an element of subjectivity in quantitative analyses, then the objectivity of its results may be in question.

For example, two quantitative researchers may make different decisions in fitting a statistical model to observed data. If there is any gray area in the correct procedures, the decisions made by both researchers may be equally justified in the absence of a prescriptive process. It may be that neither decision has bearing on the overall outcome of the model, in that either path has a negligible result on the general inference made from the data. Conversely, it could be that the decisions made in statistical analyses can alter the inference made. A small decision made in the course of data analysis may have a ripple effect, conceivably affecting whether an intervention reaches a level of statistical significance and, in turn, brings about a decision to pursue or terminate the program.

Conflicting Conclusions

Conflicting conclusions are not uncommon in social science research. Replication of research studies to confirm results is a hallmark of the scientific method. By reproducing results, the scientific community can ensure that an important result from a given study was not a fluke result and yield information about the generalizability of results to different populations or settings. Attempts at replication can also yield conflicting results, however, and lead to debate around the true results of a scientific study.

Consider, for example, previous research conducted in Florida that found the perception of post-high school economic prospects was a strong predictor of dropping out of high school (Bickel & Papagiannis, 1988). More specifically, school districts in counties with higher wages and lower unemployment had lower dropout rates than districts with fewer economic opportunities post-high school. This study was replicated in West Virginia to confirm the results and test the generalizability of the Florida findings (Bickel, 1989). In West Virginia however, the variables measuring prevailing economic climate failed to produce significant regression coefficients in predicting high school drop out rates.

Another example comes from the evaluation of the Transition Mathematics program, a curriculum intended to improve math skills in students in the 7th to 12th grades. Hedges et al. (1986) found that the program increased math achievement in students more than a comparison curriculum. Baker (1997) later replicated the study and found that students participating in Transition Mathematics performed lower on math achievement. Finally, in a follow-up study by Thompson et al. (2005), students participating in the Transition Mathematics curriculum showed no improvement over students participating in a comparable curriculum. Three studies examining the same intervention yielded three different results. From the perspective of policy makers, it then becomes unclear whether Transition Mathematics is effective and should be implemented on a wider scale.

There are several potential reasons why there was a failure to replicate the original findings in these examples and in other occurrences in social science research. For example, in the studies examining the relationship between high school dropout rates and perceived economic climate, the failure to replicate results may have been the result of differing population demographics between the two states (Bickel & Papagiannis, 1988). In the case of Transition Mathematics, differing inferences concerning the effectiveness of the curriculum may have resulted in differences in how it was implemented at the local level (Thompson et al., 2005). These are plausible and commonly referenced explanations for variations in study results. One explanation not mentioned by any author in the examples above, however, is that there may have been differences in the steps taken in data analysis; the researchers may have used differing statistical models to make inferences about the program. This explanation warrants examination to determine whether it is a source of disagreement across research findings.

Variation in Research Design

Applied social science research is a multistep process. These steps include designing the research or evaluation study, selecting a sample to conduct the study, choosing which variables to collect data on, analyzing the data for trends, and communicating the results to interested parties. This, of course, is not an exhaustive list of the steps required in a research study. Nevertheless, it does give a broad overview of the process. Importantly, errors of inference can occur any of these steps.

Research design (e.g., experiments, quasi-experiments, case studies) is a moment of consideration that greatly affects quantitative analyses and can also produce errors of inference

(Shadish et al., 2002). Design decisions can produce erroneous inferences, particularly in generalizing results of a study to various populations. For example, there has been much deliberation regarding when it is appropriate to employ quasi-experimental designs over randomized experiments and the impact of causal effect estimates (Cook et al., 2010). Likewise, there has been much consideration of how research design, sample selection, and choice of measures can affect inferences, as is evident around the discussion regarding the What Works Clearinghouse. Less frequent are examinations around the actual data analysis aspect of research, particularly in quantitative research.

Tourmen (2009) examined the choices that evaluators make in the course of their work, particularly in the evaluation design phase. She noted that an evaluative situation can be approached by any number of methodologies or prescriptive theories. The approach used by an evaluator may be influenced by the experience of the evaluator and/or the personal and evaluation theories that are adopted by the evaluator. Tourmen found that these factors affected how different evaluators responded to certain evaluative situations and, in turn, the methodological choices they made. Although Tourmen did not follow the evaluations to their final products, the decisions accrued during the process of these different evaluation approaches presumably affected the type and quality of information provided to stakeholders.

Tourmen (2009) focused her study only on the design phase of evaluation. At the other end of the research and evaluation process, there are also decisions on how to communicate the results of findings to interested stakeholders after analysis of relevant data is concluded. This is particularly true in evaluation studies, where an implicit goal is being able to communicate evaluation findings in a manner so that interested stakeholders can use that information (Alkin et al., 2006). Indeed, many scholars have argued that it is the responsibility of evaluators to present findings to policy makers or program developers in a manner that is conducive to those stakeholder groups using those findings (Fife, 1992). There are decisions to be made at this point of the research process as well. Jones and Mitchell (1990) noted, for example, that there may be a tradeoff between "precise, technically correct" reports and those that are more "user-friendly" and more widely disseminated (p. 449).

Overall, the choices made in designing a research or evaluation study can have implications for the type of information collected and the validity of findings down the line. In addition, decisions made regarding how to communicate findings affect how they are used in a policy arena. It is within the realm of reason that some sources of varying inferences are also the result of the decisions that take place at another point during the research—that is, during the course of data analysis.

Subjectivity in Data Analysis

Some scholars have questioned the assumption of statistical analysis as objective information. For example, Kritzer (1996) described quantitative analyses as an interpretative process, drawing on the context and experiences of the analyst to draw conclusions from the data. Krtizer continued to describe quantitative data analysis as an art form, and statistics as a performance similar to an actor or an orchestra. This echoes sentiments shared by Cronbach (1982), who stated, "designing an evaluative investigation is an art." These descriptions of the statistical data analysis process move away from objective sentiments that are implicit in the evidence-based practices movement.

Some decisions concerning the application of statistical methods in the social sciences are more clear than others. For example, when comparing three groups for mean differences on a given outcome, it is much more desirable to conduct a one-way analysis of variance as opposed to multiple independent samples t-tests to control the family-wise Type I error rate. The motivation behind the statistical model used in this contrived example is dictated by straightforward statistical theory that shows an inflation in error rates when multiple tests are conducted. However, some decisions are not as easily governed by statistical theory, or they may use rules that are more arbitrary. It is possible that other cues, such as the experiences or theoretical foundations of an analyst, may affect the decisions made during the course of quantitative analysis, which in turn may affect the resulting statistical model and the inferences drawn from that model.

Tekwe et al. (2004) examined whether differences in model specifications provided inconsistent results in the context of value-added model scores to evaluate teacher effectiveness. One of the original value-added models, the Tennessee Value-Added Assessment System, was criticized for not having included student demographic variables as covariates in the model (Ballou et al., 2004). Ballou et al. (2004) argued that it was unnecessary to control for student background characteristics because baseline scores for students implicitly captured student demographics. Tekwe et al. examined this assumption as well as other differences in model specification. Ultimately, these scholars compared four different value-added models and found differences in the value-added measures being produced and the inferences regarding specific teachers' effectiveness. McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004) corroborated this finding by discovering that not controlling for student socioeconomic status greatly influenced value-added scores, particularly in schools with heterogeneous student populations.

Clearly, at least in the context of teacher value-added model scores, how a statistical model is specified can affect the final results and the inferences derived from these results. It is argued here that this may also be an explanation for failing to replicate results in social science research. In Tekwe et al. (2004), the decision to include student demographics as a covariate, among other model specifications, affected the results of teachers' value-added scores. It is conceivable that such decisions in the course of model specification may also affect inferences about whether an intervention to improve math scores is effective or whether there is a relationship between dropping out of high school and perceived economic climate, as well as other matters of concern to public policy. The following section describes the different decisions that are typically made during the course of quantitative data analysis.

Decisions in Model Specification

Statistical analysis is governed by a prescriptive process intended to increase objectivity in such analyses. However, in practical real world settings, the logic and rules of statistical applications begin to be encroached by situations that may not always be governed by the prescriptive rules that guide quantitative analysis. Although there are several rules that provide structure to any quantitative analysis, there exist many instances in the course of data analysis where the deductive rules of statistical analysis do not perfectly apply. It is not uncommon to hear a professor in statistics advising students that although the computer software will calculate all the sample statistics, standard errors, and inferential tests requested by the user, it is still up to the user to correctly set up the statistical analysis and interpret the results. That is, there is a still a human component to quantitative analysis.

There are rules that govern the general course of data analysis, such as which statistical models to apply in which situations. However, these rules do not comprehensively prescribe the course of data analysis. Whereas some aspects of data analysis can be described as linear or straightforward, there is still plenty of freedom for quantitative analysts to diverge in the course of examining the same data set. For example, a correction might be needed so that the data being analyzed conform to the assumptions of a statistical model; this might mean transforming the data to meet normality assumptions or eliminating outliers so that they do not overly influence parameter estimates. There are several such decision points in the course of a typical data analysis where the researcher must decide how to proceed.

Leamer (1978) labeled these decision points in applying statistical models as "specification searches." His work centered on six different types of specification searches. A *hypothesis testing search* refers to setting up the statistical model to test the correct null hypothesis. For example, a statistical test might test a null hypothesis that there is no growth over time, or that growth trends are not related to differences in group assignment. A *data selection search* refers to how subgroups are sliced from an overall analysis, whether to split groups into separate analyses, as in analyzing minorities and White students separately, or analyzing the entire sample as a whole. As Leamer described such searches, "the same theoretical hypothesis underlies all three specifications: the one estimated with all the data and the pair estimated with the subsets. The specifications differ in their choice of data sets" (p. 7). A *proxy variable search* concerns the process of how constructs are operationalized. The analyst may make choices to use different measures or operationalizations that purportedly assess identical constructs. *Post data model construction* is a "process of revising the underlying theory in response to the data evidence" (p. 8). Practically, it is a process whereby covariates are entered into the model in an exploratory fashion to see whether the covariate improves theory and/or the statistical model" (p. 8). In an *interpretive search*, restrictions are placed on the data, such as holding constant certain relationships between variables, in hopes that certain estimates will be improved. Finally, a *simplification search* refers to the process of eliminating variables to create a more parsimonious model. Leamer presented these specification searches as a means of identifying the volatile nature of quantitative data analysis and helping analysts discover ways in which bias may creep into their analysis. In future work, Leamer acknowledges that "no one has ever designed an experiment that is free of bias, and no one can." (p. 33, 1983). Leamer developed these specification searches to help overcome this human limitation.

Leamer (1978) presented these specification searches as neutral and mind-free in terms of how they are actually specified. Although Leamer allowed for the role of the analyst's prior experience in helping to guide specification searches, he did not include a thorough explanation of how these experiences might influence model specifications. Rather, the specification searches were presented descriptively. That is, he intended only to describe the types of specification searches. However, what is needed is an understanding of *how* these specification searches are made and the cognitive and metacognitive processes required to make such decisions. To what extent does experience play a role in specification searches? What types of experiences are needed to conduct specification searches? How does the quantitative analyst draw on previous experiences to help make specification decisions? These types of questions are left unanswered by Leamer's work, but we must begin to explore them in order to ascertain the extent to which final statistical model specifications might be influenced by different prior experiences.

Simmons, Nelson, and Simonsohn (2011) also coined a phrase to describe the decisions made in quantitative data analysis. Specifically, they described *researcher degrees of freedom* as the flexibility that is inherent in most data analyses. However, the authors did not delineate the types of researcher degrees of freedom the same way that Leamer (1978) created a typology of specification searches. As described by Simmons et al:

> In the course of collecting and analyzing data, researchers have many decisions to make: Should more data be collected? Should some observations be excluded? Which conditions should be combined and which ones compared? Which control variables should be considered? Should specific measures be combined or transformed or both? (p. 1359).

There is some overlap between Simmons et al.'s (2011) researcher degrees of freedom and Leamer's (1978) specification searches. For example, asking what conditions to combine and what conditions to compare might be analogous to Leamer's data selection search. The consideration of what control variables to use is an example of post model data construction or a simplification search. Finally, combining or transforming variables might describe specific techniques in an interpretative search. However, Simmons et al. pointed out additional analytic decisions that were not covered in Leamer. For example, they also asked about the appropriate sample size in an analysis and how outliers are treated in a model.

As Simmons et al. (2011) made clear, the subjective decisions that lead to self-affirming conclusions is "not the by-product of malicious intent" (p. 1359). The authors continued, arguing that these behaviors are the result of two factors: "(a) ambiguity in how best to make these decisions and (b) the researcher's desire to find a statistically significant result" (p. 1359). It is presumed in the current research that the former is more culpable than the latter. In fact, it is assumed that decisions made by quantitative researchers during data analysis are implicit or subconscious, and that no underlying purpose guides researchers to act in a certain way.

To be sure, there exist many cases of questionable research and analysis practices that serve self-interest rather than public knowledge. A recent publicized example includes the case of social psychologist Diederik Stapel, who was found to have fabricated data for over a decade and in over 50 publications, including in student dissertation work that Stapel supervised (Levelt et al., 2012). On a more institutional level, House (2008) pointed to the many ways that researchers who study prescription drugs, funded by large pharmaceutical corporations, are able to manipulate their research trials to produce more favorable results.

In cases of explicit fraud, it is clear that quantitative analysts are motivated by self-serving interests in the form of publications and profit, and no study into these cognitive processes is warranted. These extreme cases are not considered in the current research project. Rather, it is presumed here that the only motivation for researchers is to uncover true relationships in the world. Therefore, the current study makes no attempts to uncover explicit self-serving biases in researchers; only the value-free decisions that may ultimately affect research findings are under examination.

Updating the Leamer Typology

The framework of Leamer (1978) was introduced nearly three decades ago. There have since been many methodological advancements and technological developments since that time that warrant an update on Leamer's initial framework. Several of Leamer's specification searches are temporally universal and apply to contemporary quantitative data analyses, including those that were observed in the current analysis. However there are also other limitations that prevent the framework from fully applying to the currently collected data. For example, hierarchical linear modeling (Raudenbush & Bryk, 2002) is much more prevalent in current educational research than it was during the time Leamer was developing his framework and thus does not incorporate decision types that are relevant for these models. There have also been new developments in handling missing observations since the conceptualization of Leamer's framework. Specifically, multiple imputation methods have been developed that offer a more rigorous analysis of data with missing observations (Rubin, 1987). None of Leamer's specification searches directly address decisions related to missing observations. Presumably, since these techniques simply did not exist during Leamer's time, analysts did not have the option to utilize them and did not have to make decisions about how missing observations would be handled.

Leamer's (1978) framework includes one particular specification search that he called "postmodel data construction," what he described as "the process of revising the underlying theory in response to the data evidence" (p. 10). However, the argument might be made that this is not a discrete decision during data analysis, but rather the analytic process as a whole. Researchers may tweak and refine models as new information comes along or as new thinking on a topic occurs. The question remains around how exactly the model is improved upon and what alternate decision might be made to improve the model. This specification search by Leamer does not specify what kind of adjustments are to be made. It also follows that different researchers would make varying decisions to try to improve their models. These unanswered questions with Leamer's framework suggest that perhaps an updated framework is in order. Leamer attempted to help practitioners of econometrics and statistics to become better analysts and produce better inferences. Leamer laid the foundation for exploring and describing this topic and the current research study aims to update and expand upon his work.

Models of Decision Making

The process of decision making has been a subject of considerable study in a wide variety of fields and disciplines, ranging from nursing, teaching, and designing evaluations to more specific tasks such as grading essays and conducting an orchestra (Barkaoui, 2011; Bergee, 2005; Borko & Livingston, 1989; Stough & Palmer, 2005; Swanson et al., 1990; Tourmen, 2009; Westerman, 1991; Williams et al., 2010; Wolfe et al., 1998). Of particular interest is how individuals make judgments and decisions on a particular course of action when a problem is encountered in the performance of a specific task. A great deal of this line of research involves contrasting experts and novices in a given discipline or task and comparing cognitive processes between levels of proficiency (Chi, 2011).

The study of decision making draws from multiple disciplines, including cognitive psychology, neuroscience, computer science, organizational behavior, economics, sociology, and political science (Wang & Ruhe, 2007). Since the scholarly work on decision making derives from such disparate disciplines, there is no consensus model or framework that is unanimously endorsed by researchers in this field. There is, however, a common language shared by many scholars in decision making that describes the cognitive processes in making a decision. Wang and Ruhe (2007) described three components in this process: setting a *decision goal,* choosing between *alternative choices*, and maintaining *selection criteria*. Chi (2011) offered a similar conceptualization. According to Chi, "the representation of a well-defined problem consists of its elements, all the permissible operators that can operate on the steps of the problem, the constraints on the operators, and the goal of a problem" (p. 19).

Several models have been offered by decision making researchers to represent the decision making process. Some of these representations include mathematical models that describe rational decision making behaviors (Wang & Ruhe, 2007). Other models are graphical and help illustrate what Chi (2011) described as the decision making "space of all possible moves" (p. 19). One such illustration comes from Hastie (2001). This graphical representation is reproduced in Figure 2.1.

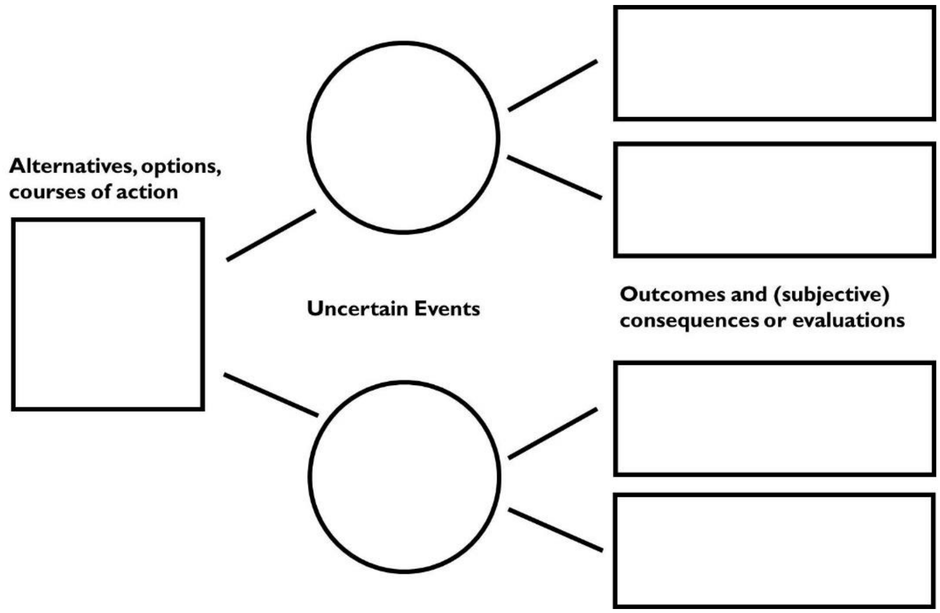


Figure 2.1. Template decision making framework model (adapted from Hastie, 2001).

Hastie's (2001) model uses a decision tree to visualize problems as a fork in the road, a point at which a decision is made about which path to follow. Similar to the conceptualization in Wang and Ruhe (2007), the Hastie model illustrates three major components of decision making: alternative courses of action, outcomes or consequences, and uncertain conditioning events. Starting from the left side of the model, decision points are reached that contain alternate paths or courses of action that are permissible. Branching from these choices are paths that lead to specific events. Each event carries some level of uncertainty, and there is never a guarantee that a decision will lead to a specified event or outcome. Therefore, each event also branches out to two different outcomes or consequences of the action being taken. Although the Hastie model contains only two paths at this juncture, it is possible for multiple courses of action to be permissible, thus expanding the model exponentially. In other words, the model can be adapted to accommodate the many possible consequences that can arise when decisions are made in real world settings.

According to the Hastie (2001) model and others (Wang & Ruhe, 2007; Chi, 2011), the outcomes and consequences on the right side of the model drive the decisions to be made at the starting point on the left side of the model. An outcome is deemed desirable by the actor making the decision, and then the choice is made as to what course of action best leads to that desired outcome. There is some level of uncertainty in these decisions and their eventual outcomes, as not all decisions have a corresponding one-to-one relationship with any particular outcome. Several courses of action may lead to the same outcome, for example. Or, a specific course of action may have several consequences, some of which may be unintended and/or undesirable. Therefore, there is a judgment component where the actor assesses and evaluates the events and outcomes that are likely to occur after a specific course of action.

The Hastie (2001) model is general enough to be adapted to many different contexts, including decisions made during statistical analysis. For example, consider how to handle an outlier that is found during data analysis. For the sake of simplicity, suppose there are two courses of action available: The analyst may either delete the outlying observation or decide to leave it in the data set. Each decision has a pair of corresponding potential outcomes. If the outlying observation is deleted, an estimated coefficient estimate might be less subject to bias or a statistical model might have a better fit to the observed data. Or, deleting the outlier might result in a loss in statistical power, with no marginal benefit in terms of producing a more unbiased inference. Alternately, keeping the observation might enhance the external validity of the analysis. If the analysis in question pertains to student test scores, for example, then the outlying observation might correspond to a gifted student who scored unusually high relative to his or her peers. Or, the outlying observation might represent an individual who scored high on a survey designed to measure certain psychopathic traits. In each of these cases, deleting the outlier would restrict the analysis and resulting inferences to those that fall in the “normal” range of scores.

The analyst must make the decision about whether to keep or delete this observation. This choice is made in conjunction with the perceived outcomes that are most desired by that analyst. If there is a greater emphasis on technical accuracy, then deleting the outlier may be the more appropriate course of action. However, if greater external validity and generalizability are desired, then keeping the observation might be the more appropriate choice. Or, there might yet be other outcomes and consequences desired by the analyst that influence the decision. This statistical analysis example and possible decision making options are adapted to the Hastie (2001) model in Figure 2.2.

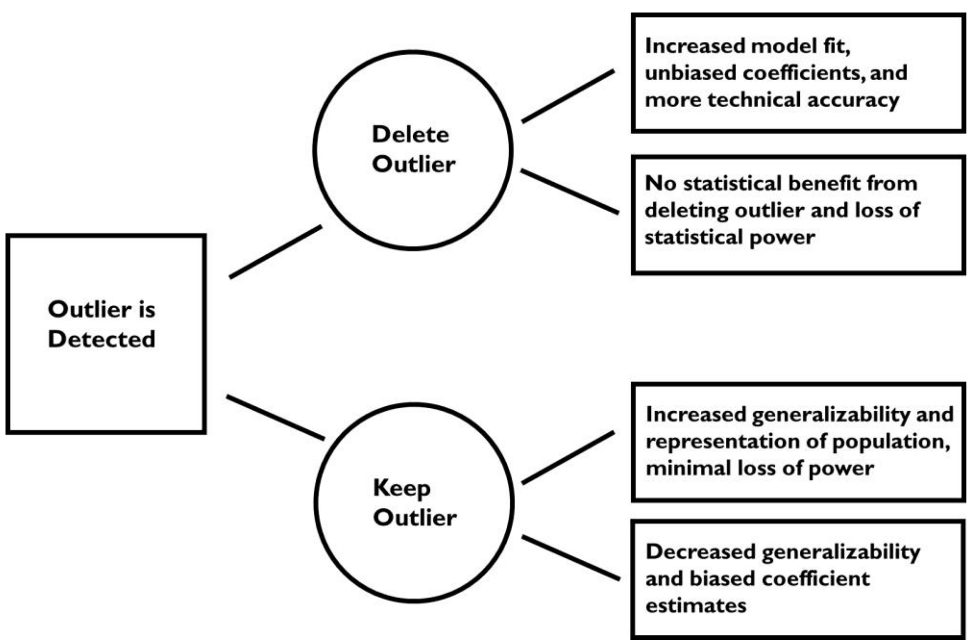


Figure 2.2. Decision making model for deleting an outlier.

The Hastie (2001) model simplifies the decisions made by quantitative analysts by presenting decisions and outcomes as a choice between only two alternatives. In the above example, the only two options were to delete the outlier or to keep the observation. In practice, alternative courses of action are also permissible, such as using a logarithmic transformation on the variable so that the outlying observation is not an outlier in log units, or dichotomizing the variable by way of a median split or some other cutoff and specifying an alternative model. There are also more than two potential outcomes from each course of action. Additionally, outcomes were presented in terms of statistical outcomes of power, model fit, and unbiased estimators. However, other potential outcomes may exist in the perception of the quality of the analysis. The model without the outlier may be perceived to be more technically accurate methodologically, for example, or the model without the outlier may appear to be less credible by certain constituents and stakeholders who may wonder why the gifted student or the person with greater symptomatology is not considered in the data set. Decisions in practical settings are typically more complex than the Hastie would model suggest.

Limitations of the Hastie Model

Hastie (2001) provided a helpful model to think about how judgment and decision making have been conceptualized by academic research. The model is not without its limitations, however, particularly when applied to decision making while analyzing quantitative data. One essential problem is the subjective nature in which outcomes are evaluated. There are two layers to this subjectivity: the outcomes themselves are subjective to the degree that each is desirable and in the propensity of each to occur, since probability is typically evaluated before any action or decision has occurred. There is, in many circumstances, very little previous information available that provides insight into the likelihood of a particular outcome occurring after a decision has been made. Since the outcomes are evaluated before a decision is made, the decision hinges on the accuracy of the decision maker in assessing these subjective outcomes.

The question then arises as to the nature of this subjectivity, or how these subjective evaluations are influenced. More specifically, little is known about how evaluation of outcomes is influenced by external sources, particularly in the context of quantitative data analysis. As we will see, researchers often do not make decisions in a vacuum and are often influenced by external sources of information that narrow the decision space. Thus, the present analysis pursues an adaptation of the Hastie (2001) framework that gives credence to how these subjectivities are influenced and the various sources of influence within the setting of quantitative data analyses.

In Hastie's (2001) framework on decision making, outcomes are treated as exogenous rather than endogenous. That is, according to Hastie, subjective outcomes are evaluated to assess the preferred course of action in a decision tree. However, there are antecedents that influence how these subjective outcomes are produced. Subjective outcomes do not solely affect how decisions are made; rather, there are external factors that impact how they are produced. Since the outcomes are not innate to the decision maker, it follows that there are considerable spheres of influence that prescribe how they are constructed. Subjective outcomes are malleable, and it is important to understand how they are affected to fully understand the cognitive process of decision making.

Studying Experts

Studying experts in specific domains has been one method of looking at how individuals make decisions in specific fields of study. The logic of studying experts in their respective fields is attractive to researchers because it allows insights into how people become proficient in the skills necessary for their respective professions. An expert is generally defined as an individual who has gained superior knowledge and skills in a particular domain (Stough & Palmer, 2005;

Glaser & Chi, 1988). Implicit in the study of expertise is that skill proficiency is not an innate quality of an exceptional individual, but rather something that is developed over time and ultimately shapes the cognitive processes necessary to perform a certain task (Chi, 2011). Thus, novices have the capacity to become experts in a given discipline or skill, and studying how experts perform tasks may elucidate how an individual reaches expert status.

Studying experts also can reveal competencies that are necessary for a profession, in hopes of increasing the workforce in a given profession where there is a dearth of labor. For example, Stough and Palmer's (2003) study on the qualities of expert special education teachers was motivated in part by the need to staff special education classrooms with competent personnel. Not surprisingly, studies of expert teachers have increased over the past decade, particularly as public policy has increasingly focused on teachers and classroom processes.

These studies have yielded many insights that help distinguish the cognitive processes of expert teachers from those of novice teachers. Expert teachers certainly possess a greater knowledge base than novice teachers. However, not only do expert teachers have a greater knowledge base than novice teachers, they also differ in how this information is organized and the type of heuristics they use to solve problems in the classroom (Borko & Livingston, 1989; Swanson et al., 1990). This in turn affects how expert and novice teachers differ in their interpretation and perception of events in the classroom (Westerman, 1991).

Chi (2011) postulated that in any given discipline, several qualities exist that separate experts from novices—for example, the ability to better understand a problem and know the possible constraints to a problem and the permissible steps that can lead to a solution. This is described in the context of solving algebraic equations or playing chess, where certainly the rules of algebra and the game of choice define the permissible moves allowed. Statistical analysis also has structural rules that help define the space of possible moves. Experts in statistics might be able to identify the space of possible moves better than those taking their first course sequence in statistics, in part because they simply know more of the methodologies available for the situation at hand. In a decision about how to handle missing observations in a data set, for example, multiple imputation techniques might be known to the expert but not to the novice.

Another quality that Chi (2011) described as differentiating experts and novices goes beyond just the capacity for knowledge to suggest that experts and novices have differences in how that knowledge is structured and represented. Chase and Simon (1973) showed that in the context of a chess game, expert and novices alike were similar in their recall of the positioning of chess pieces if the pieces were placed randomly on the board. However, if they were placed in the context of meaningful chess plays, then experts were able to recall the positioning to a much greater extent. Thus, experts were able to "chunk" domain-specific information into more meaningful units than novices; their knowledge was structured in a more salient way.

Chi (2011) also showed differences in how experts and novices represent knowledge and are able to focus on different aspects of the information being provided to them. Chi illustrated this in the context of physics problems given to college students. When given identical problems, novice students tended to focus on surface properties or concrete elements of the problem, such as what type of physical objects were represented. Meanwhile, advanced students were able to focus on the deeper elements, such as the physical law that was underlying the problem. Put another way, novice students focused on superficial aspects of a problem whereas experts represented the knowledge in ways that were relevant to their specific domain.

There is evidence that these expert and novice differences are also apparent when moved away from controlled settings of chess games and physics problems and into real world issues, such as those encountered in professional settings. In her study comparing expert and novice teachers' decisions in lesson planning, Westerman (2001) found that expert teachers were able to structure the lesson and curriculum from the perspective of the student and what they have had previous exposure to. They were able to relate the current lesson and scaffold new ideas onto what students had previously learned. Novice teachers tended to focus primarily on learning the objectives of the current lesson, without making connections to what students might already know. It appears that expert teachers were able to integrate, or "chunk" elements of the new curriculum with what students already knew and what they had been exposed to in prior lessons; novice teachers could not "chunk" information in this way and talked about the finite and individual aspects of the curriculum.

Tourmen (2009) also found differences in the way expert and novice evaluators assessed an evaluative situation. Tourmen's results suggest that new evaluators tended to design evaluations in a manner that was more methodologically-oriented and focused on how the methods learned could be implemented step-by-step. More experienced evaluators, however, tended to be more flexible in prescribing certain methods, and instead focused on the utility of their methods with respect to the evaluative decision at hand. Experienced evaluators were also able to identify more contextual factors in the evaluation situation, whereas novice evaluators tended to focus on the informational needs of the evaluand and the methods suited to fulfill those needs, without regard to context of the setting.

It is unclear, however, whether these constructs translate to the specific domain of statistical analysis. More experienced quantitative analysts may have more knowledge than their less experienced peers, but is this knowledge structured and represented differently? One goal of the current study was to identify how knowledge is structured and represented in statistical modeling.

Metacognition

There is little question that statistical analysis is a cognitive task. That is, statistical analysis demands that the researcher think about the task and use certain mental processes as they are performing it. However, less is known about the metacognitive processes that are used during statistical analyses. Although there are some inconsistencies in the definition of the construct, metacognition generally refers to the awareness of one's cognition and the regulation of cognitive processes (Veenman et al., 2006). More colloquially, metacognition may refer to a "higher-order cognition about cognition" (p. 5). As opposed to skills, which refer to the procedural knowledge used to solve a problem, metacognition refers to the declarative knowledge used to solve problems. For example, the skills used in statistical analysis may involve knowledge of computerized software and the steps required to run a statistical model. Metacognition, or declarative knowledge, in statistical analysis involves an understanding of how and why these steps take place.

Metacognition in the performance of any task may be operationalized by being able to describe the steps required to complete the task, including defining the goal of the task, planning a course of action to complete it, and evaluating whether the plan is being executed correctly and/or working to complete the task. Since it is a higher order level of cognition, metacognition draws heavily on cognition (Veenman et al., 2006). As Veenman et al. (2006) noted, "in terms of metacognitive skills, one cannot engage in planning without carrying out cognitive activities, such as generating problem solving steps and sequencing those steps" (p. 5). With regards to statistical analyses, cognition is required to identify and apply a certain strategy used in the course of the analysis, or to be able to interpret the results of statistical tests, whereas metacognition refers to knowing why the strategy might be effective in a particular circumstance, or being able to troubleshoot whether the strategy was effective or not.

Metacognition is described as a self-regulating feedback loop of cognition (Schraw, 1998). Individuals monitor their cognitive strategies and whether they are working to achieve a specified goal. These concepts can be applied to the domain of decision making. In the Hastie (2001) model of decision making, specific courses of action may lead to specified outcomes and consequences that are anticipated by the actor. However, this is only a probabilistic model and actions may not always lead to specific outcomes and may actually produce unintended consequences. A metacognitive feedback loop might monitor the linkage between actions and perceived outcomes and assess whether these linkages are indeed accurate. If an unintended consequence is produced from a specific action, the actor may reflect on why he or she believed in these linkages and adjust their internal schemas accordingly.

For example, a potential job seeker may have to choose between two potential job prospects. The job seeker may evaluate both options and ultimately decide on one position based on the perceived outcomes that the selected position would be more satisfying and in a less stressful environment. If these outcomes are not realized, the employee would engage in a metacognitive process and evaluate his or her cognitive process to determine why it was originally believed that the job would be satisfying or less stressful and what lead to these errors in judgment. Future career decisions would be informed by what the individual learned through that metacognitive process.

One of the goals of this study was to discover the metacognitive processes that occur during the course of statistical analysis. The cognition and skills required to perform statistical tasks can be found in traditional statistics textbooks; less is known about the extent to which metacognition is used to regulate these cognitive tasks in the domain of statistical analysis. An extension of the proposed research attempts to understand the variability in metacognition as it pertains to researchers performing statistical analyses. If there are differences in metacognition between researchers—either across disciplines or between levels of expertise—this could very well explain the differences in the decisions made by different researchers during the course of data analysis.

CHAPTER 3

RESEARCH METHODS

In the tradition of anthropology and the sociology of science, the current research study posits that the quality of science and research is improved by casting a lens onto the scientific process itself (Atran, 1998; Latour & Woolgar, 1979). However, in addition to examining the social factors that affect scientific endeavors, the present study is also concerned with the personal and psychological factors that may affect research findings. In that vein, this study employed exploratory research methods to provide a descriptive account of the cognitive processes of quantitative researchers while performing quantitative data analyses. As such, the study relied primarily on qualitative research methods.

It has been argued so far that quantitative data analysis is a complex cognitive process. Despite rigid prescriptive rules advanced by statisticians on how to proceed during different conjunctures that typify the data analysis process, there exist many situations that are not governed by standard statistical procedures. It is thus left to the quantitative researcher to determine how to proceed in such scenarios. A qualitative study is warranted to examine these cognitive processes. This approach will allow us to move beyond reductionist and formulaic models of decision making during the data analysis process. In line with this, the study examined the following primary research questions:

(1) What types of decisions are made by quantitative researchers during the data analysis process?

(2) What are the cognitive and metacognitive processes of quantitative researchers during the course of data analysis?

(3) What types of information do quantitative analysts rely on to solve problems and make decisions during data analysis?

(4) To what extent do these cognitive and metacognitive processes differ across disciplines?

The first three research questions aimed to offer a descriptive cognitive and metacognitive account of decision making during quantitative data analysis. The intended end product was a model that describes the cognition that occurs during this process. In contrast, the fourth research question attempted to explain the variation, if any, of cognitive processes in quantitative researchers. It is possible that not all quantitative researchers make the same decisions and employ the same underlying cognitive process. It is easily reasoned that the type of training or the general norms of a specific scientific discipline might influence how quantitative researchers make certain types of decisions. Thus, it is important to understand how well the cognitive model derived from this research can be applied to quantitative researchers in different fields.

Participants

When this study was initially conceived, a total of nine quantitative analysts were targeted to provide sufficient data to explore the cognitive processes during data analysis. I anticipated using purposeful sampling techniques to select these quantitative analysts from a variety of academic disciplines, including economics, sociology, public policy, and education. Scholars in each of these disciplines are interested in educational issues, albeit from differing perspectives and methodologies that may affect how they treat ambiguous situations in quantitative analyses. Analysts from other disciplines, including psychology and anthropology, were also considered. The goal was to include a diverse sample of participants from a variety of disciplines that deal with educational issues.

Furthermore, the sampling pool was initially defined to be researchers who had used and published articles using the Trends in International Mathematics and Science Survey (TIMSS). This data set was selected because of its prolific use in educational studies in a variety of disciplines. By using an internationally collected data source, issues of data quality would ostensibly be avoided and the study would be able to focus on data analysis decisions that were not affected by poor data quality.

Authors of TIMSS publications and those presenting research on TIMSS findings at national conferences were contacted for potential participation in the study. Approximately 60 researchers were contacted through this method. Many researchers expressed interest. However, after a short period using this recruitment strategy, it became clear that this approach was problematic because there was often a lag between when researchers conducted analyses using TIMSS and when they were able to publish or present their research. Many researchers who expressed interest in participation had moved on to other research projects that did not necessarily involve TIMSS.

For these reasons, the sampling pool was broadened to include researchers using any state, national, or international data set collected by any governmental or foundational entity. With this change, more researchers were able to participate using data sets they were currently analyzing. Furthermore, by using data collected by government agencies and foundations, data quality was maintained. Once the sampling pool was opened up, additional participants were recruited through word of mouth and purposeful sampling. Principles of theoretical sampling (Glaser & Strauss, 1967) were employed to recruit participants in later stages of the research process, after initial theory development and during the course of initial coding. Specifically, participants with an academic background in economics were purposefully sampled to generate a more comprehensive sample across disciplines.

Ultimately, eight participants were included in the study over seven cases; two participants in the sample worked as a team on the same research project. Although this fell short of the initial goal at the outset, they represented a variety of backgrounds such as teacher education, economics, and policy analysis. Participants varied in the amount of time that they participated in the study. Some participants' involvement consisted of the minimum level of participation described below. Other participants had more immersive relationships that consisted of followup emails and hallway conversations that supplemented the standard data collection procedures.

Study participants varied in their level of expertise. Some research participants were graduate students applying quantitative methodologies that they had recently learned. Other participants were tenured or former professors with a long line of research publications. This diversity in expertise allowed for insights to be made around how decision making in quantitative data analyses might be affected by experience.

The following section briefly summarizes each participant in the study and the analysis that was observed. Distinguishing details have been altered to guard against identifying the specific projects and researchers. For the most part, constructs and analytical decisions were not changed, but specific variables may be disguised. The context behind each decision and details pertaining to how decisions were resolved were preserved as best possible.

Brian

Brian was a recently-hired assistant professor at a research university in the Southern United States. He participated in the summer after his first full year as a tenure track faculty member in the school of education at his university. Brian had completed his doctoral studies in teacher education the year prior to his faculty position. It was his first faculty position, although he was a part-time instructor at his graduate university in the Western United States. Brian's main research interests involved teacher education and teacher quality. He was particularly interested in measuring teacher practice and improving instructional quality, especially in mathematics.

Brian was relatively new to quantitative methodology, as his graduate program emphasized more qualitative methods. At the suggestion of his advisor, Brian began to pursue more quantitative methods midway through his program in order to make broader claims about his research. Brian had to take coursework outside of his department and learn a lot of quantitative methodology on his own. Eventually he also attended short trainings at conferences and national education agencies. Brian has since presented his findings at national conferences and published his research with colleagues in peer-reviewed journals.

Brian's analysis was focused on the relationship between teaching practices and student achievement. He was using a large-scale comparative education database to conduct his analysis. In particular, Brian was interested in comparing different models of instructional practice and how these practices may differ across countries and cultures. His analysis used regression-based methods to understand the relationship between teacher instructional practices and student achievement.

Shay

Shay was a graduate student in education at a Midwestern research university. She was a math teacher before she entered her current graduate program to study teacher education. She described her methodology training as primarily qualitative. Her research interests included factors that predict student achievement, particularly in mathematics education. Her research also incorporated a comparative education component, as she was interested in how high-achieving students look in other countries.

Shay described herself as a novice in quantitative methodology; although she took coursework in statistics in her undergraduate and graduate institutions, she described them as basic. It was not until she took a course in hierarchical linear modeling, which she described as "the most amazing class," that she decided to pursue her research interests with quantitative methods. Shay was a graduate student instructor in her program, but the course she taught was mainly focused on content and teaching practices, rather than statistics or methods.

Shay's analysis was focused on students' background factors and demographic characteristics that were related to high achievement in math. She was using a large-scale comparative database for her analysis and incorporating regression-based methods for her analysis. Although her current research project was only focused on the United States, she hoped to eventually expand her research to understand high achieving students across multiple countries.

John & Stephen

John was an associate professor in education at a university on the West Coast. Although he had been at his current institution for only a few years, he was a professor at a university in the Southern United States prior to his current position, also in a school of education. His teaching load consisted of courses in comparative education and education in developing countries. John had been involved in numerous comparative education studies, conducting research in Africa, Asia, and Latin America on topics related to educational resources and the impact of international policy. John had published extensively on these topics in several research journals and book chapters.

John received his doctoral degree in comparative education at a research university on the West Coast. His program and training involved consulting with numerous education economists, although he did not necessarily consider himself an economist. Similarly, John had taken numerous research methodology courses and had been involved in extensive research projects, but he did not consider himself a methodologist.

Stephen was an international graduate student at the university where John was a professor. Although John was not Stephen's advisor, they had collaborated on a couple of projects in comparative education. Stephen had a primary interest in comparative education when he first attended graduate school in the United States. He had attended a pervious graduate program in the Midwest, where he first starting taking courses in applied research and quantitative methods. Although he was relatively new to statistics and quantitative methods, Stephen took a liking to the content and attended as many statistics courses as he could.

The school of education that Stephen was attending was not focused on quantitative methodology, but he was able to take courses in other departments, such as economics and psychology, to supplement the curriculum. The university had recently been hiring professors with strong quantitative backgrounds, and Stephen was able to leverage his position as a graduate student at the university to learn more diverse quantitative methodologies from the recently hired professors. Together, John and Stephen had presented their research at several national and international education conferences.

Their research used a large-scale comparative database to understand rural and urban differences in student achievement across multiple countries. Their project was, in part, an expansion of a previous study that was being updated with more current data. Their research also aimed to evaluate a specific education model designed to reduce urban and rural differences in student achievement in a specific country. The analysis included propensity score matching and hierarchical linear modeling. For the most part, during the course of their analysis, Stephen conducted the analysis and presented interim results to John. The two discussed these interim results together and collaborated on potential modifications and next directions in the analysis.

Peter

Peter was an assistant professor for a research university on the West Coast in the school of education. His primary research interests included measures of school and teacher effectiveness. Of particular interest to Peter was examining the distribution of quality schools and teachers across students of various socioeconomic statuses, and looking at school and teacher effectiveness as a means of closing the achievement gap for students from underserved communities.

Peter attended an undergraduate institution at a major university in the Midwest, but received his doctorate at a research university on the West Coast, where he specialized in quantitative methodology. Peter had been an assistant professor at his current university since 2007. Before that he also taught for a university in the Southern United States. He was a math and science high school teacher before he obtained his doctorate. Peter had been an author on numerous journal articles and book chapters on educational topics throughout his career.

Peter's analysis used a national longitudinal database that tracked students' transitions from high school into adulthood. The focus of his analysis were demographic characteristics of students' high schools and school cultures, a construct termed "organizational habitus" (McDonough, 1997). He was particularly interested in how these factors played into students' decisions to enter selective colleges. Specifically, his analysis utilized a mediation model to examine whether these constructs mediated the relationship between socioeconomic status and college choice.

Tracy

Tracy was a graduate student at a research university on the West Coast. She was near completion of her program in economics, just needing her dissertation to complete her degree. As an undergraduate math and economics major, Tracy had extensive experience with quantitative methodology. Her university was especially rigorous in economics, as Tracy claimed that she had completed a full year more of econometrics than would typically be required in other economics programs.

Tracy was an international student, originally from Southeast Asia. While attending school in the United States, she worked as a research assistant and completed her first publication. Under the tutelage of her advisor, she analyzed data from a randomized controlled trial in India. Tracy categorized herself more as a methodologist than as someone with specific content expertise, but her research interests included labor economics and racial migration.

Tracy's analysis examined the relationship between worker training programs and worker outcomes in STEM industries using a nationally representative data set about the labor market. Specifically, she was tasked with examining whether worker training programs affect wages differently for those in STEM fields versus those not in STEM fields. She was using a differences in differences model for her analysis, a traditional econometric model that is commonly used in her field.

Eric

Eric was a senior research analyst for a state department of education on the West Coast. Before working for the state, Eric was a research associate for a not-for-profit research institution also on the West Coast. His primary responsibilities included looking at factors of school quality and effectiveness and factors related to academic achievement. Eric received his doctorate degree in economics with an emphasis in educational policy studies at a university in the Southern United States.

Eric had prior teaching experience as an adjunct professor teaching statistics and quantitative methodology. He had done analyses on several projects and evaluations and prided himself on being current in quantitative methods despite it being many years since he earned his doctoral degree. Eric had published just under a dozen articles and presented at numerous professional conferences throughout his career.

Eric was conducting an analysis examining the relationship between family involvement with state assistance programs and educational outcomes for these children. More specifically, he was interested in family trajectories and how involvement in public assistance or the child welfare system affected school achievement outcomes, including high school graduation and college attendance.

Natalie

Natalie had been working as a research analyst for the department of education for a state on the West Coast for the past two years, and for a different agency in the same state for the previous six years. Prior to that, she was an analyst for a research institute on the East Coast, where she received her graduate degree in education. Natalie’s primary research interests included school reform and educational policy. She completed her doctoral studies as the charter school movement was beginning to unfold during the 1990s, but her more recent work primarily looked at measures of school quality.

Natalie described her graduate program as mainly policy-oriented. Although she took several courses in quantitative methodology during her program, her expertise in statistics had mainly been derived from the work she did during her career after receiving her doctorate. Natalie primarily used regression-based methods in her work, including hierarchical modeling. A great deal of Natalie’s work was communicating findings to leaders at the state and local level.

Natalie’s analysis was focused on seeing the effects of a community college workforce development program on state unemployment rates. Her analysis looked at several sites across the state and was longitudinal, as she examined unemployment rates over time, comparing counties that had implemented the workforce development program to those that did not. She was using a time-series analysis in her work.

A summary of the participants is included in Table 3.1. The table lists the title, sector, and discipline of each participant. Also included is a brief description of the analysis that was observed as well as the data that were collected for each analysis.

Table 3.1

Summary of participant demographics, analysis, and data collection sources

Pseudonym	Title	Sector	Discipline	Analysis	Data Sources
Brian	Assistant Professor	University	Teacher Education	Regression analysis of teacher instructional practices related to student achievement	Pre Interview Analysis Observation Post Interview
Shay	Graduate Student	University	Teacher Education	Regression analysis of student demographic characteristics related to high student achievement	Pre Interview Analysis Observation Post Interview
John & Stephen	Associate Professor & Graduate Student	University	Comparative Education	Propensity score matching and hierarchical linear model analysis of rural versus urban differences in student achievement	Pre Interview Analysis Observation Weekly Meeting Notes Post Interview
Peter	Assistant Professor	University	Educational Research Methods	Mediation analysis on the effects of socioeconomic status on college selectivity through school organizational habitus	Pre Interview Analysis Observation Post Interview
Tracy	Graduate Student	University	Labor Economics	Regression and differences in differences model on the relationship of worker training programs on labor outcomes	Pre Interview Analysis Observation Post Interview
Eric	Senior Research Analyst	State Agency	Education Economist	Regression based analysis on the effects of involvement on government assistance programs on school achievement outcomes	Pre interview Analysis Observation Email Correspondence Post Interview
Natalie	Research Analyst	State Agency	Policy Analyst (Education)	Time series analysis on the effects of a community college workforce development program on unemployment	Pre Interview Analysis Observation Email Correspondence Findings Presentation Post Interview

Materials

The primary materials used in the study were the data sets used by the participants. These data sets included international comparative data sets, national labor market data, and administrative data collected at the state level. Participants used their own computer and statistical software for their data analysis in the environment of their choosing. Any other software normally used by quantitative researchers, such as word processing or spreadsheet software, was also allowed. Participants had their data analysis process recorded according to the procedure described below.

Procedure

Participants were contacted and screened for inclusion in the study. Those who elected to participate initially underwent a three-step data collection process. First, a pre-task semi-structured interview was conducted with each researcher. This interview provided context for the analysis. The questions shed light on the demographic characteristics of the quantitative analysts, their familiarity with the data sets they were using, previous training in quantitative analyses, and other contextual factors that may have influenced their decision making in quantitative data analysis. The interview was audio recorded with the consent of the participants and later transcribed for coding. The protocol for this interview is included in Appendix A.

Second, a think-aloud procedure was conducted after the pre-task interview, during quantitative analyses of the data set of the participants' choosing. That is, observations were conducted on data analyses that research participants would have performed as part of their original research agenda. Research questions were screened to ensure that it was appropriate for inclusion into the study, and that it was focused enough so that the data collection process was not too burdensome on the participant. The think-aloud technique allowed insight into the cognitive processes of each subject while he or she was performing certain tasks.

This research method has been used previously in studies of decision making (Barkaoui, 2010; Bergee, 2005; Lundgren-Laine & Salantera, 2009; van Someren et al., 1994). Van Someren et al. (1994) argued that the process of verbalization concurrent with task performance yields insight into the cognitive behavior and reveals the information that is stored in a person's working memory at any given moment while performing a task. Others have criticized the technique because verbalizations may only reveal cognitive processes derived from information stored in a person's working memory and therefore deeper cognitive processes may not be revealed. Furthermore, the act of verbalizing thoughts may inhibit the cognitive process required to complete the task on hand (Nielson, Clemmensen, & Yssing, 2002).

It is suggested that interaction between researcher and participant be minimized, so as not to interfere with any cognitive process (Lyle, 2003; Nielson et al., 2003). Therefore, the think-aloud portions were not probed by the interviewer, to make the completion of the data analysis as naturalistic as possible. The only interactions anticipated between the researcher and participant during the think-aloud task were reminders to participants to verbalize their thoughts if there was a sustained period with no verbalizations offered by the participant (Ericsson & Simon, 1993).

Although the intention was for the think-aloud task to include a minimal amount of interaction between the data analyst and the observer, in practice some participants had difficulty with *not* interacting with the observer. Therefore, some observations of data analysis featured more dialogue between analyst and observer than anticipated. As the research progressed, it became clear that dialogue between the analyst and the observer was a fruitful method for discovering the cognitive processes of data analysts. Although the observer never initiated any verbal prompts directly to the data analyst, questions from the data analyst to the observer were addressed and were not discouraged.

The think-aloud task for each participant was recorded using Screencast-o-matic, an online software that comes in free or subscription-based versions. The software can screen capture the user's computer while superimposing audio and video via a webcam into the screen capture. Screencast-o-matic is online based, circumventing any issues with having to install software on participants' computers; any computer with a Java-enabled internet browser is able to utilize the software. This allowed participants to use their own computers with their software in the setting of their choosing. Concurrent to the think-aloud task being performed by the participant, observation field notes were also collected. The observations provided another source of data, and helped guide the post-interview, as described below.

After the think-aloud procedure was complete, a third data collection procedure was utilized. A post-task interview was conducted with each participant following completion of analysis on their data set. The interviews were semi-structured and followed a format similar to the initial pre-task interview. The purpose of this second interview was to allow the participants to reflect on their task performance and to think aloud. Questions on the post-task interview were intended to elicit responses detailing the extent to which the analytic processes of the participants during the task performance replicated what they might have gone through during an actual data analysis of their own choosing. This was in part to examine the ecological validity of the task performance as well as to further understand the cognitive processes that may not have been apparent during the task performance. The post-task interview also specifically asked about certain analytic procedures that were observed during task performance, if appropriate. Participants had the chance to elaborate on their decisions during the task and to reflect on other approaches that might have been considered. The post-task interview was adaptive to the specific data analysis procedures and strategies used by the participating quantitative analyst during the task performance. Therefore, the protocol included questions that emerged while observing the think-aloud. A copy of the semi-structured interview protocol used after data analysis is included in Appendix B.

The post-task interview was structured around capturing three specific domains related to task performance: procedural, cognitive, and metacognitive. The procedural domain of the interview asked researchers to recount the steps and actions used to complete the data analysis. The cognitive domain pressed on why these steps were undertaken and how participants made the decisions necessary for completion of the data analysis task. Whereas the procedural domain spoke strictly to the sequential steps used in the data analysis, the analytic domain attempted to ascertain the thinking patterns behind these steps. Finally, the metacognitive domain probed deeper into the thinking process used in data analysis and asked participants to reflect on their mental processes.

For example, some data analyses used the transformation of quantitative variables to make distributions that were more normal, as assumed by certain statistical models. The procedural domain captured whether a transformation was used and, if so, which transformation. The cognitive domain probed on why specific model specifications were used, or how preliminary analyses, such as examining histograms or descriptive statistics, informed the data analysis process. The metacognitive domain moved beyond these mechanistic processes and examined the higher order thinking that was involved. Special attention was given to how participants monitor and evaluate the validity of their data analysis.

The post-task interviews were conducted as soon as possible after the think-aloud to preserve recall of the task performance. Breaks were permitted after the think-aloud to allow participants to decompress from a cognitively demanding task. For every case, all three steps in the data collection process were conducted on the same day. The total amount of time for these three steps varied across individuals, but ranged from approximately 6-8 hours per participant.

The data collection strategies above yielded several rich sources of data: two interviews (one conducted prior to task performance and another just after); a video recording of the task performance during the think-aloud protocol; and field observation notes taken during the think-aloud. In many instances, the data were supplemented by other data sources. Most commonly, casual conversations during and after any of the data collection strategies elucidated the thinking taking place during data analysis and were thus captured and analyzed whenever possible. Some participants also sent emails following the think-aloud session, most notably when they discovered a mistake that needed to be fixed in the model specified. Furthermore, John and Stephen held weekly meetings where decisions about model specification were discussed. These meetings were especially helpful in understanding the rationale behind their decisions, and they were also recorded, transcribed, and analyzed as data. Another participant presented her initial findings to an informal group of researchers to discuss the preliminary results. That informal presentation and a subsequent interview were also recorded, coded, and analyzed.

Data Analysis

All interviews, observations, meeting notes, and video recordings were transcribed. Transcriptions were completed as soon as possible after each interview, data analysis session, or meeting. Preliminary coding and analyses were conducted concurrently with data collection since there was often a lag between recruitment of participants. This allowed for more targeted interviewing and observation in later participants as themes emerged.

Analyses were conducted using principles adapted from grounded theory (Glaser & Strauss, 1967; Strauss & Corbin, 1990). Practically, this meant that codes, concepts, and hypotheses were derived inductively from the data rather than relying on preconceived notions of how constructs should be related. Grounded theory procedures allow for the development of theory that is "grounded" on collected data. The current research project was an ideal setting for grounded theory procedures as there have been few prior research studies on this particular topic and the primary objective of this project was to generate theory.

Video recordings and transcriptions were uploaded and coded using NVIVO version 10.0. Coding of interview and video data was originally informed by the model specification search categories outlined by Leamer (1978) and the Hastie (2001) model of decision making. Potential preliminary codes included the type of specification search, components of the decision making process, and domains of cognitive and metacognitive processes. Preliminary analysis of the data coincided with transcription of interviews and coding of the data. Videos were coded with specific detail to decision points encountered during the course of data analysis.

Although coding schemes were partially informed by the frameworks mentioned above, the analysis was also driven by themes that emerged from the data, using inductive strategies to build theories of decision making in statistical analysis. Indeed, preliminary coding frameworks derived from Leamer (1978) and Hastie (2011) were eventually scrapped in favor of new coding frameworks that were emergent from the data, in accordance with qualitative data analysis methodology adapted from grounded theory procedures (Glaser & Strauss, 1967; Strauss & Corbin, 1990). Categories of codes were adapted to accommodate these themes. Codes were analyzed for similarities and differences across study participants, with particular attention to differences associated with academic discipline.

As noted above, data collection, transcription, and data analysis were not discrete points during the course of the study. Analysis began before data collection was complete. As such, theoretical models began to emerge as new data were collected and new participants were observed and interviewed. This allowed for a constant comparison of new cases that allowed for corroboration of the emerging theoretical models.

The final codebook used to code and analyze all interview transcripts, documents, and observation notes is included in Appendix C.

CHAPTER 4

DECISION MAKING IN QUANTITATIVE DATA ANALYSIS

This chapter will explore the decision making process during the course of quantitative data analysis as seen in observations of study participants conducting their research. The first part of the chapter will describe the types of decision made by study participants while conducting data analysis. Then, factors affecting the decision making process in quantitative data analysis are discussed.

As described previously, Leamer’s (1978) framework for describing specification searches in quantitative data analysis is not without limitations in contemporary analysis. Advancements in methodology and software have made Leamer’s framework outdated. His original framework does not account for these advancements and therefore does not completely describe the types of decision made in current-day quantitative data analysis.

Aside from advancements made in quantitative methodology, Leamer was decidedly a Bayesian, even devoting the second chapter in *Specification Searches* (1978) to serve as a primer on Bayesian analysis[1]. None of the analyses observed during the current research project used Bayesian techniques as a foundation for the analysis. Rather, frequentist methods were primarily used. As such, the language and terminology used by Leamer did not fully apply. Leamer even conceded this fact, stating that “it is difficult to find non-Bayesian language that can make such a statement less ambiguous, but in the Bayesian language, hypothesis-testing searches make use of alternative specifications that are assigned positive subjective prior probability” (p. 9). His hypothesis-testing specification search is essentially the reworking of hypotheses about the

[1] Bayesian analysis refers to a school of statistical analyses that attempts find the probability of a given hypothesis, starting with a subjective probability distribution and then adjusting those probabilities based on observed data. In contrast, frequentist methods attempt to find the probability of observed data, given a hypothesis of the researcher. Interested readers may refer to Leamer (1978) or Gelman, Carlin, Stern & Rubin (2004) for more information.

relationships between observed variables being tested through the use of alternative Bayesian prior probability distributions and constraints on variables. Indeed, none of the analyses that were observed for this research project used subjective probabilities, let alone ones that were modified and retested. For these reasons, this chapter will discuss the decisions made during quantitative data analysis using examples and observations grounded in the data rather than applying a deductive approach using the Leamer framework.

Chi's (2011) essay on expertise provides a useful metaphor when thinking about decision making. Chi postulated that experts, when compared to novices, have a greater understanding of the space of possible moves and are able to search this space with more proficiency when encountering a problem. In statistical data analysis, decisions of how a model is specified abound, and in each of these scenarios the researcher must decide how to handle the situation. That is, the researcher must determine what the space of possible moves is and how to navigate that space. The decision space is vast but finite, and its borders are defined by the sensible approaches that reasonable researchers would consider acceptable for a particular research project. This metaphor will be relied upon throughout the presentation of the current research findings.

Types of Decisions in Quantitative Data Analysis

Selection

Selection refers to choosing among an array of possible variables and cases to be included in a statistical model. Selection of variables involves choosing among available covariates in a model to increase statistical power or precision in a model. Examples include whether to use student demographics as a covariate in a value added model assessing teacher quality (Tekwe et al., 2004; McCaffrey et al., 2004). Selection of cases, on the other hand, refers to any exclusion criteria that might occur when deciding whether to include certain cases in an analysis. For example, if an analysis is focused on low income students, exclusion criteria might include whether the student is receiving free or reduced lunch or if the child's family's annual income is below a certain threshold. Deciding which criteria to use and which thresholds to establish are considered selection decisions.

In some instances, a selection of covariates to include in an analysis might parallel with a simplification search in Leamer (1978). Although none of the analyses observed for this project exhibited this process, it is conceivable that some quantitative analysts might choose to eliminate a vector of non-significant predictor variables in a regression model to increase parsimony and interpretability. These instances would constitute a simplification search because extraneous variables are being eliminated to produce a more useful model.

One of the study participants, Brian, examined the relationship between teaching practices and student achievement across countries, but not all countries available in the data set were included in the analysis. How Brian justified his selection of countries out of the overall pool available would constitute a selection decision across cases. Similarly, Brian chose only a subset of teachers' instructional practices to include in the model, rather than using all that were available. How Brian chose which indicators of instructional practices to include would constitute an example of a selection decision across variables.

Structure

The structure of the analysis refers to whether variables or cases are kept structurally intact or are broken down into subcomponents for the sake of analysis. When dealing with continuous variables, a researcher may decide to dichotomize or categorize a variable into nominal groups based on some criteria. Likewise, subgroup analyses may be conducted on categories of cases to see if there are differential effects on certain segments of the population. Leamer (1978) referred to this type of decision as a data selection search. His hypothetical example included an analysis for the overall sample as well as two separate analyses for participants in the North and South. He stated that "the same theoretical hypothesis underlies all three specifications: the one estimated with all the data and the pair with estimated subsets. The specification occurs in the choice of data sets" (p. 7).

As Shay explored factors related to high academic achievement in math, two analytical decisions made during the analysis related to structural decisions. One predictor variable included in the model indicated the average time the student spent on homework per week. This variable was a continuous variable, but was bimodal rather than continuous. Shay decided to split the variable into intervals given the distribution of the data. This would be classified as a structural decision across variables since the original variable was broken down into categories for purposes of the analysis.

Shay's data set provided variables measuring students' performance in mathematics; these scores were standardized and therefore continuous. Shay had to define a subgroup of the overall population as high achieving for her analysis. This serves as an example of a structural decision made across cases, as Shay had to classify her observations into categories.

Construction

Construction decisions pertain to how variables and cases may be combined in certain analyses. Variables are often combined into scales or indices to represent a certain construct of interest. Choosing which variables to combine and how to combine them are decisions of construction. Also, subgroups may be aggregated to provide statistical power or parsimonious results as desired by the researcher. For example, ethnic categories in a demographic analysis might be combined into a general “other” category if few cases are observed in certain ethnic categories. When applied to decisions over cases, construction decisions can be thought of as the opposite of structure decisions; construction decisions aggregate groups into a broader category while structure decisions break down general classifications into more specific subgroups.

John and Stephen created a socioeconomic index from the observed variables in their data set, several of which might have been possible indicators for socioeconomic status. How they decided which variables to use and the appropriate method to combine these variables into a single measure would constitute a structure decision across variables. Likewise, Natalie’s analysis of the effects of a community college workforce development program provides an example of construction over cases. Natalie had to group cases depending on their employment status during each data collection point in a longitudinal study. Several employment outcome categories were available in Natalie’s data set, including participants who were fully employed, employed part time, seasonally employed, or self-employed. How Natalie decided to group these cases together in her analysis describes constructions across cases or observations.

Model Assumptions

Decisions related to model assumptions refer to choices that reflect the distributional assumptions of certain statistical models. For example, in many regression-based methods, outcome and predictor variables are assumed to be normally distributed. In cases where the variable does not meet this assumption, a transformation, such as a logarithmic or exponential transformation, might be used to better meet model assumptions. Another common assumption is that the data are free of influential outliers. Decisions can be made by the researcher around how to detect and define an outlier, as well as how to handle outlying cases.

Although decisions related to model assumptions do not completely map onto Leamer's (1978) framework, there are some analogous specification searches. Transformations used to meet model assumptions can be thought of as a hypothesis-testing search. Essentially, the analysis shifts from finding the relationship between variables on their original scale into finding the relationship between variables on their transformed scales. The detection and possible elimination of outliers can also be thought of as a data selection search, in that the choice of data set being analyzed should be free of influential outliers.

Tracy, for example, used a transformation of her dependent variable in her analysis of the impacts of workforce training on labor outcomes in STEM fields. Workers' wages were one operationalization of labor outcomes in her analysis. However, wages often have a skewed distribution, which would not be optimal for many statistical analyses that assume a normal distribution. How Tracy decided that a transformation was necessary, as well as which transformation to use, offer examples of decisions related to model assumptions across variables.

Natalie's analysis of counties implementing a workforce development program provides an example of a decision related to model assumptions across cases. Given the geographic diversity of the state, there were different rates of unemployment across counties—her specific interest in her analysis. One county had an unusually high unemployment rate, even when compared to other more rural counties in the state. How Natalie handled this outlying observation constituted a decision related to model assumptions, because she had to consider the implications of the outlier on her statistical model, which assumed the absence of outlying observations.

Generalization

Finally, decisions around generalization reflect how the results of the specific data analysis might allow the researcher to generalize results to other settings or increase external validity. Often, researchers are interested in relationships between broad theoretical constructs that are represented by one or more observable variables. These proxy variables may or may not represent the construct of interest. Decisions of generalization pertaining to variables include selection of proxy variables that validly represent the construct of interest or how missing observations are handled in a data set. Missing cases often decrease external validity because observations might not be missing at random, and the pattern of missing cases might indicate bias (Pigott, 2001; Little & Rubin, 2002).

Leamer's (1978) framework also included a proxy search, which he used to describe the process of finding a variable or set of variables to represent an underlying hypothesis. Leamer acknowledged that many observed variables are measured with error. Analysts that find little or no relationship between hypothesized variables might not be able to distinguish between a poor proxy variable and a true non-existent relationship between constructs.

When examining the impact of school organizational habitus on students' college attendance and selectivity, Peter had to find proxy variables to represent this high-level construct. He selected a data set for analysis that was longitudinal so that high school factors could be linked to students attending college. That is, his choice of data set was motivated more by the ability to track students over time than how well it represented his construct of interest. His choice to use these proxy variables as operationalizations of his construct would be classified as a decision related to generalization because it required that he accept an assumption that the measure validly represented the construct of interest.

Decisions related to generalization across cases involve handling missing variables for cases of individuals. In all analyses, researchers had to worry about how missing data would be handled. Missing observations are ubiquitous in social science research and working with secondary data sets does not offer an escape from missing observations. All researchers who participated in this study were aware of the missing observations in their data sets and used different strategies to handle the problem.

Interactions Between Decision Types

These decision types are not necessarily discrete. A decision made earlier in an analysis may affect the inputs for another decision further down the road, and a decision in one category can impact a different decision in another category. John and Stephen, for example, were compiling variables to include in an index of socioeconomic status, a construction decision type across variables as described above. They looked through a list of demographic variables available in their data set, two of which were mother's education and father's education. There was a high degree of missing data, however, even when looking at these two variables. Stephen described a strategy where "I've excluded, for instance, mother's education and father's education separately because there's a lot more missing data across the two than if you just take the highest between the two… Missing data ranges about 18% per variable, but when you do it across, it's about 32%."

This problem was further compounded when more variables were included in the index for socioeconomic status. Stephen discovered that when the full list of desired variables was included in the index for socioeconomic status, "it will be over 53% listwise deletion, and only about 46% remaining." John asked whether the high degree of missingness was associated with any variable, to which Stephen responded, "there is not a single variable that is a big outlier, where it has 60% or 70% missing; they're all in the 10 to 15% range. The more (variables) you add, the more listwise deletion. So it's just the likelihood of listwise deletion gets increased, even though the single variables don't have that much difference in terms of missing data themselves."

This above vignette displays the intersectionality of decisions in quantitative data analysis. The first step in creating their socioeconomic index was a selection decision regarding the correct covariates to include in the index. This process was directly related to decisions about generalization, as missing observations for these variables were compounded as more variables were selected. Since there is no single direct measure of socioeconomic status, John and Stephen created a proxy variable by indexing these covariates into a single measure. How this proxy index variable was constructed (i.e., the technical aspect of how these variables were combined) was another decision that had to be considered after the covariates were selected. Finally, these decisions had to be juggled with missing data causing some concern with a high percentage of cases being omitted. So the decision of which variables to include in an index was directly related to which cases to include in constructing the index because of the compounded missingness when more variables were included.

Factors Affecting Decisions in Data Analysis

The previous section explored the types of decisions that are made during quantitative data analysis. The focus of the remainder of this chapter is to expand on those initial concepts and begin discussing *how* decisions are made and the factors that influence the decision making process. Put another way, rather than describing the space of possible moves, this chapter will begin to elucidate how quantitative researchers navigate through this space. Chi (2011) postulated that experts have a better understanding of this space when compared to novices.

However, as we will see, the decision space is more heavily restricted by external sources other than just subject matter expertise.

Social Influences on Data Analysis

The quantitative researcher is often perceived as a professor in an office, sitting in front of a computer, crunching numbers through statistical software. The analyst is mechanistically following the prescribed procedures of quantitative analysis to produce results that will either support or refute a given hypothesis. Although this conception is not entirely false, and indeed much work occurs in office-like settings, it fails to capture the collective aspect that can drive quantitative analysis.

In fact, analysis is often shaped by the social group of a researcher's peers. A researcher may validate his or her decisions by seeking the advice of a mentor or colleague when unsure about the correct course of action during data analysis. While conducting their analyses, the researchers in this study often wondered aloud how others might handle a particular situation. John and Stephen, for example, often speculated how a revered colleague would handle their missing data; John rhetorically asked, "What would [name omitted] do here? She must have written that in her article." Shay also wondered how an advisor would have proceeded in her analysis: "I am trying to imagine what [my advisor] is saying to me right now."

Other examples provide evidence of how a peer group can more directly influence the decisions made by a researcher during data analysis. Consider Natalie, a research associate at a state department of education. Natalie was working on an analysis assessing the impact of a college workforce program on employment outcomes across the state where she was employed. Specifically, she was working on the cross-site analysis of the program that was implemented across community colleges and corresponding counties in the state. During the course of her data analysis, Natalie had to make decisions of selection to decide how unemployment would be measured. Her decision was not how to measure unemployment, as the unemployment rate was provided in a single, interpretable number from the state's employment department and this number had been provided monthly for many years, longer than the timeline of Natalie's study. Rather, the decision Natalie faced was how to use the monthly unemployment figures or, more specifically, which monthly figures to use. Natalie succinctly captured the decision at hand:

> (These) counties are very seasonal in their employment. Everyone comes in the summer and anytime else it's dead. We will need to account for the changes in seasons. But I don't have October through December numbers for the last year. I can not use last year, or I can not use October through December for all years. Let me think. What to do?

With these short utterances, Natalie described the fork in the road in the analytical decision space. It is also clear that Natalie simplified the decision space into a dichotomous choice. At the heart of this particular decision were the missing observations in the data set. In the most recent calendar year of the analysis, the last three months of unemployment data had not yet been made available by the employment department. Natalie reasoned that seasonal changes in monthly unemployment numbers made these missing observations problematic. Thus, omitting the last three months for the most recent year would yield an inaccurate representation of the average unemployment rate for the year when aggregated across months. Since this strategy would have produced a biased estimate in the most recent year, it was not a permissible move in the analytical decision space, and Natalie did not consider it. Instead, she saw two permissible moves in the decision space—completely deleting the most recent year of observations or eliminating three months of observations for each year under consideration.

If the analysis was stopped at this point in time, the Hastie model (2001) would provide an accurate representation of how Natalie made her decision to handle the missing observations. She simplified the decision space into two dichotomous choices and furthermore subjectively assessed the consequences of each of the two choices. If Natalie chose to omit the last year of observations, the consequences of that decision would be an incomplete picture of the effects of the workforce development program, as the truncated timeline would not allow the program enough time to develop and show positive effects. On the other hand, removing three months out of each year would give an admittedly biased account of the unemployment rates for that year when factoring in seasonal employment trends throughout the year.

After some internal deliberation, Natalie decided to move forward with nine months of unemployment rates representing the entire year of unemployment. "We handle the seasonality of unemployment, make every year equivalent, and [are] also able to keep last year, which we are mainly interested in," Natalie reasoned. She subjectively weighted the inclusion of the last year of data, truncated as it may have been, as more favorable over omission of the last year of data. She implicitly favored losing a year's worth of data as less desirable than losing three months of data annually. Mechanistically, Hastie (2001) provided an explanatory model that at this point accurately described Natalie's decision to select which observed variables to use to represent her construct of unemployment. At this point in time, the analysis was not complete, however, and the subsequent actions and ultimate reversal of Natalie's decision reveal the limitations of the model.

A few weeks after completing her analysis, Natalie presented her initial findings to a dozen or so researchers in an informal setting. These researchers came from not only from the department of education where Natalie worked, but also from other agencies in state government, including those at the state's employment agency. They gathered in a small conference room where Natalie had set up a projector in the front of the room. She began her presentation by explaining the purpose of her study and the goals of the workforce development program before eventually reaching the methodology portion of her presentation. She explained how three months were omitted from each year of the analysis in order to be able to include the last year of data and to control for seasonality effects.

After Natalie finished her presentation, she solicited questions and input from the audience. It was not long before someone in the room asked the first question. "So you didn't include October through December in your analysis?" asked someone near the back of the room. He continued, "That would take away the part of the year where unemployment would be the highest in some areas." Before Natalie could form a response, another person chimed in with another viewpoint on the topic: "It seems that you are not really adjusting for seasonal fluctuations so much as ignoring it." The criticisms kept coming until the question and answer session evolved into a discussion about how to appropriately represent unemployment rates.

After approximately 15 minutes, the three researchers who were involved in the discussion reached the conclusion that omitting October through December from all years was not an appropriate method for controlling for seasonal effects of unemployment and was therefore not a permissible move in the analytic decision space. The researchers agreed that Natalie should extend the timeline of the research project, so that the last three months of unemployment data could be released for the most recent year in the analysis. Natalie was mostly silent during this time, but receptive as she listened to the conversation and nodded in agreement to the points being made. In an interview a week after her presentation, Natalie informed me that she would follow the advice of her colleagues and delay the results of her findings until more data were available. She revealed that she had brought up the issue of seasonal unemployment to the employment research analyst, but "it wasn't until presenting the results to the group that anyone had any concern."

Natalie's struggle reveals a great deal about the nature of the data analysis process. She initially made a decision about how to accurately represent unemployment rates; she subjectively weighed the consequences and outcomes of her decision and decided on the best approach. That decision was overruled by the input of her peers, however, and the consensus was that Natalie's methodology was no longer tenable. Instead, her colleagues suggested an approach that was not under consideration when Natalie first made her decision. Through a social process, the decision space was expanded and reduced at the same time—expanded in the sense that another option in the decision space opened up that was not previously considered, and reduced in that this new option was now the only decision that was collectively accepted by these researchers and therefore the only permissible move in the decision space.

Natalie's story of how unemployment was represented and specified in her statistical model was not an isolated incident among the researchers I observed. Recall the process of Stephen and John, the graduate student and mentor studying rural and urban differences in achievement in a comparative education setting. Their process involved weekly meetings to discuss the previous week of analysis and to troubleshoot any issues encountered. This space was also used for analytical decisions anticipated during the next round of analyses, which would occur in the coming week. The two met face to face to discuss the analysis and anticipate criticisms that might arise. This was contrary to Natalie's process—she conducted the bulk of the analysis by herself, only to receive guidance afterward, when she presented her findings. Whatever the actual process, both accounts highlight that analysis happens in social settings—the office or the presentation room—as much as it does behind a computer in front of a statistical software package or spreadsheet.

Effects of Peer Review

Peer review is one mechanism through which social influences can affect decision making in quantitative data analysis. The process is intended to ensure high quality research through the blind review of an expert panel. In addition to maintaining a high level of rigor in journal submissions, however, the current study also offers evidence that the peer review process may influence researchers during the course of data analysis.

One salient example comes from John and Stephen in their analysis of rural and urban differences in student achievement. One major consideration was how to handle missing observations. Their analysis involved multiple segments, first using propensity score matching to identify appropriate comparison groups and then a multilevel model to compare group differences. The pair had already decided to use multiple imputation for the multilevel modeling, but were deciding whether it would be appropriate to use multiple imputation techniques with propensity score matching. John stated, “If we’re using multiple imputation for the [multilevel] model, some reviewer is going to say, ‘Why didn’t you use it for the propensity score matching?’ if we don’t use it for the propensity score matching.” In part because of the anticipated pushback from a potential peer reviewer, John and Stephen decided that the missing data had to be handled through multiple imputation, not only for the multilevel modeling, but also for the propensity score matching phase.

Peter, the tenured professor, also agreed about the influence of peer reviewers, stating that, “you always have to be aware of what a journal reviewer will say, whether or not their comments are fair.” Although there was no evidence that peer reviewers were on Peter’s mind while he was conducting his analysis of the effects of school organizational culture on student college attendance, it was evident through his testimony that there had been times in the past where a peer reviewer's comments had shaped his analysis.

The influence of peer review was not evident across all settings, however. When asked about pressures revolving around the peer review process, Natalie, a researcher at a state agency, responded, "I don't need to worry about that. The only people I have to convince are those that are actually going to use the findings…to make decisions." Later on she added, "I have to care about being accurate and having good information, and not about the agenda of a peer reviewer." To Natalie, the potential concerns of a journal referee were not a consideration that governed her decision making.

Clearly, those who are affected by the process of peer review are those who submit manuscripts for publication. Researchers in non-academic settings where there are fewer pressures to publish in peer-reviewed journals are arguably less susceptible to a process that does not affect their career standing. Indeed, among the sample being observed, researchers in state government positions did not share concerns about peer reviewers. In addition to Natalie's eloquence in dismissing peer review as a consideration, Eric simply added that he "left that world behind," referencing a previous position at an academic institution, where presumably publications were more valued and thus peer reviewers were more salient.

Software Capabilities

Software capabilities and limitations can also be an external influence that restricts the decision space for researchers. This limitation can manifest itself in a few different ways—either the user does not have the technical knowledge of how to use a particular function in the software being used, or the software itself is not capable of handling certain methodological functions. For example, Brian, in his analysis of effective teacher practices, provided a cogent example of the latter scenario, where software limitations restricted the use of methodological practices that might have otherwise been considered.

Brian was using data from an international comparative data set to make inferences about the impact of various teaching styles on student achievement scores in several different countries. The data were naturally nested at various levels, providing a structured data set that could be used in a hierarchical linear model. The student-level achievement data were nested within teacher data, which were in turn nested within school and country data. Not only was a hierarchical linear model a method for accounting for the intra-class correlation between students within specific classrooms or schools, it is the preferred way to account for a violation of independence of observations in contemporary data analysis (Raudenbush & Bryk, 2002).

Although these considerations could have steered Brian towards some type of hierarchical model in his very first analytical decision, he chose to analyze his data using an ordinary least squares regression model. After the analysis was complete, I asked Brian why he did not choose a hierarchical linear model that would have accounted for the nested structure of the data. Brian responded, "I had thought about it, but the software I was using does not allow me to work in an HLM model."

Brian was using a specialized statistical software that was primarily designed to work with the data set that he was analyzing. The software provided a user interface that allowed him to choose which variables to include in the model as well as what type of model to include. The program then produced syntax that could be copied and pasted into SPSS or another comprehensive statistical software package. The usability of the software came at a cost to flexibility and functionality, however. Among the options of type of statistical model to use, a hierarchical linear model was not among them. In this instance, software limitations greatly reduced the decision space that Brian was allowed to work with.

Eric provides another example of the limitations of software restricting the decision space. In Eric's case, the limitations came when he was making a decision related to representation across cases, in that he had to choose an appropriate methodology to handle the presence of missing data. When I observed Eric conducting his analysis, missing data were an afterthought. He conducted many frequencies and tabulations to explore the data, but seldom was the frequency of missing observations mentioned in browsing the descriptive statistics. He then conducted more advanced statistical modeling after these exploratory analyses; in these models, there was no accounting for the missing observations and therefore missing data were handled through listwise deletion, the default setting in the statistical software being used. After the analysis, I asked Eric about other options to handle missing data. He explained:

> It's certainly a consideration to use multiple imputation for the missing data. The problem is that [my software] can't do that, at least not the way it's configured. You would need to get an additional add-on package. I've asked [the state] about getting it. The thing is, is that this is the way it has been done before, using only complete cases. I have little motivation to push too hard on this because it will make the new analysis inconsistent with the old ones, and I will have to explain that.

Eric's testimony provides a direct example of how some software limits the choices that he was able to make during data analysis. Because the software that Eric was using to conduct his data analysis could not perform more advanced methodologies such as multiple imputation, Eric was resigned to using a less rigorous method for handling the missing observations.

Collective Acceptance

In the comment above, Eric alluded to another factor in opting for listwise deletion of missing observations over a more rigorous methodology. Specifically, prior to his current analysis, there had been many research projects at the agency that relied on listwise deletion, and these analyses set a precedent that Eric used to justify his decision to use a similar technique. This is a phenomenon that I label "collective acceptance."

Shay also used listwise deletion in data analysis. She did not cite software limitations or historical precedence to justify her methodological decision; rather, she simply stated that she did not know how to conduct analyses with multiple imputation, even though she knew this was preferred method of handling missing observations. To be fair, in Shay's case at least, there was a demonstrable lack of technical knowledge on how to implement multiple imputation methods. However, there may have been other drivers that contributed to the lack of technical knowledge.

Eric and Shay were not unique in their decisions to favor listwise deletion in lieu of more rigorous and accepted methods for handling missing data. Methodological literature reviews have demonstrated that listwise deletion and complete case analysis are by far the most predominant methods for handling missing data (Eekhout et al., 2012; Peugh & Enders, 2004). The prevalence of listwise deletion and complete case analysis demonstrates a collective acceptance of suboptimal methods over what is available in the data analyst's tool belt. This collective acceptance also favors certain methodologies over others that are also objectionable. Other methods exist, such as mean and regression-based imputation, but the use of these methods is not as common, and therefore cannot be used as a justification.

So, in Shay's case, the lack of technical knowledge was at least partially motivated by the collective acceptance of less rigorous methods. If the research community is generally accepting of antiquated methodologies, then there is little incentive for researchers to implement more rigorous and time consuming methods. If the majority of researchers used these techniques in their analysis, however, there would be more motivation to also use these methods. The lack of pressure from the academic community manifests itself in Eric and Shay's indifference to actively learning or obtaining software necessary for implementing these techniques.

Only one data analyst in my sample used multiple imputation to handle missing observations; the rest used complete case analysis. This is in line with the trends among researchers more broadly, as found in Eekhout et al. (2012) and Peugh and Enders (2004). As the testimony from Eric and Shay indicates, there are various reasons for reliance on an inferior methodology, from software limitations to general lack of training. However, there are two additional undeniable reasons for the lack of more rigorous methods for handling missing data: the simplicity of listwise deletion and their prevalence of these methods in the research literature.

These reasons are, of course, intertwined; the prevalence of listwise deletion is related to its ease of use. Even Stephen, the sole analyst who used multiple imputation methods, had to consult software manuals and the research literature not only to learn how to program the method in his software, but also to understand the effects of the methodology in the particular model that he was using. This required him to use a tremendous amount of time and resources investigating a single but crucial component of his analysis, rather than focusing on the overall analysis. Other researchers—both in the current research sample and in the sample studied by Eekhout et al.—have relied on methods that do not require additional time to research and can be used in default settings in most statistical software packages.

Research Precedence

Another external factor that affects the decision space is the role of previous research. When making an analytical decision, there is comfort in following in the footsteps of previous researchers who have worked with the same constructs and perhaps even analyzed an identical data set. When navigating the decision space, there may be a seemingly vast number of choices available, and emulating the analytical decisions made by someone prior provides a suitable justification for a particular decision.

When John and Stephen finalized the construction of the socioeconomic index described in the previous chapter, the pair began to think about the specification of their overall model. While they knew that a multilevel model was the suitable analytical strategy, given the nested structure of their data set, they had to provide the details for the appropriate specification of the model. Their data set included hundreds of observed variables measuring dozens of different constructs. These variables were also collected at several units of analyses, including at the student, family, school, and country level. With so many variables to choose from and multiple ways to aggregate measures to a higher unit of analysis, there was a seemingly vast number of specifications possible, even within a multilevel model framework.

However intimidating the number of options may have seemed, John and Stephen had little trouble navigating through the decision space. Instead, the pair followed the trails made by previous researchers, with the decisions these researchers made as their guiding light. While discussing the overall model, John said to Stephen, "I think we should go with this [Helms][2] paper and use a similar approach to them. Because otherwise, what do we base it on?" With that, the pair set forth on their analysis, with Helms as their guide. Several times throughout the course of the analysis, they referred back to Helms when they were presented with options about their research. Not only was the model specification similar to the analysis that was conducted in Helms, they also referred to how Helms constructed a socioeconomic status index, which was also included in the final model. In a few instances, John and Stephen did not follow the model

[2] The name of this source paper has been changed to protect the anonymity of the original research and of John and Stephen's research.

Helms used in his analysis; most of these deviations had to do with omitting a variable used in a previous iteration of the data set that was now absent from the current iteration.

Eric also took a similar approach, relying on previous research strategies to simplify the decision space. High school graduation rates may seem to be a relatively simple measure to operationalize, but there are many contextual factors that may complicate how it is measured. This is particularly true when there are several local educational agencies with slightly varied definitions and the researcher is attempting to find a harmonious definition across jurisdictions. For example, local programs available to some constituents may complicate the meaning of an on-time graduation. One urban district may have a partnership with a local community college that allows high school students to take courses as seniors; a rural district might allow flexibility for students to work on family farms and businesses. In both instances, high school completion is chronologically in the fifth year but technically on time. This makes comparison difficult at the state level, when a researcher is attempting to aggregate and compare graduation rates across jurisdictions that have different rules.

Eric was undeterred by the potentially problematic situation, however, as a trail through the decision space had already been made by those who encountered the same problem in prior analyses. He proceeded by aggregating high school graduation across local jurisdictions without any hesitation. It was not until later, when I asked about his methodology, that Eric offered his explanation: “Other people from different states might have a problem with this, but in [this state], that’s how it’s always been defined in other reports.” Later he added, “It would be confusing to some people if this report used a different way to calculate it than the way it’s always been.” Eric justified his operationalization of the construct by citing a research precedence in the agency where he was employed. He also appealed to the consistency of measurement as a means to not confuse users of the report. The decision space contained several different ways that high school graduation could have been defined, but the available options for Eric's decision were limited by the work of others who had previously conducted similar analyses using a particular methodology.

Previous research need not be in the exact same topic area for it to be influential in current analyses. In both examples described above, the previous studies that guided John, Stephen, and Eric were not exact replications of the research questions these analysts were investigating. They did, however, use similar data sets and measure similar constructs, even if the exact focus of the analysis was not the same. This was enough for the researchers to follow the leads of these previous researchers. So, a decision used in previous research need only be tangentially related to a current analysis for it to be potentially influential.

Secondary Data Limitations

The present research project only examined quantitative analysts who were working with secondary data sets that they did not personally collect or design. These data sets were typically collected on general topics and not specifically for particular research questions. It was common for researchers to choose a data set that fit their needs for a large representative sample, but that also measured some aspects of the constructs that they were considering. For example, the researchers in my sample chose comparative education data sets so they could compare educational practices or student achievement across countries, or longitudinal data sets so that events could be tracked over time. However, even when the data set captured most of the researcher's needs on a given topic, the nature of the secondary data set and its ability to serve many researchers may have meant that it was limited in some areas as the researchers tried to examine specific research topics.

For example, Peter was researching the relationship between school habitus and its mediating effects between socioeconomic status and students entering selective colleges. He chose a data set that was longitudinal primarily so that he would be able to see students' trajectories as they completed high school and moved on to either higher education or the workforce. His data set included some school-level variables, particularly those related to the school organizational culture that he was interested in. However, since Peter did not design the data collection instrument, he was not in control of the variables being collected. Instead, he had to rely on the secondary data to accurately and validly measure school habitus.

During his analysis, Peter stated, "It's one of those constructs that are a little mushy, difficult to put your hands on. But it's very attractive. I'm sure there'll be criticism in terms of how it's measured." Peter acknowledged that the construct validity of the variables he was using to measure school habitus may not have been sufficient. Since Peter was not the primary data collector, however, he was not able to refine the measures to more validly represent his construct of interest. He was dependent on the secondary data source and how proprietors of the data set collected their variables.

Other researchers provided instances where researchers lamented the absence of a particular variable or questioned the quality of a variable being collected. Brian was looking over the variables that indicated teacher instructional practices in his comparative education data set. He verbally listed these variables individually, trying to determine which were more indicative of a more traditional teacher-centered instructional practice and which were more indicative of a constructivist student-centered instructional practice. Brian started to become critical with the data set when he found that there were not as many variables indicating these varying instructional practices as he would have liked:

> So what I would expect more is that they tried to maybe add more variables on both sides. Try to map teachers' teaching strategy because, still, the variables are not enough to do factor analysis. Try to let the teachers answer and try to express their type of teaching by answering multiple variables. You don't let researchers conduct exploratory factor analysis, they need to include more on both sides. This is still not enough.

Brian would have liked to have done more with his secondary data set, but the limited number of variables restricted the analyses that he was able to perform. It was apparent through Brian's frustrations that he would have had more variables collected had he commissioned the data collection himself, even going on to list some examples of questions that he might have included if given the opportunity. However, the nature of secondary data analysis means that Brian had to settle for what was available to him.

Stephen and John also shared the frustration of a secondary data set. Recall that their model followed closely along with a model that was used previously by another researcher. As the pair tried to emulate this statistical model, they encountered a problem in attempting to recreate a socioeconomic index. The variables used in the previous report did not exactly align with what was available in the most recent iteration of the data set. Later in the analysis, they expressed more frustration as they noted that a critical variable was undefined. There were two variables that indicated if a school was rural or urban, and the frequencies did not agree. Perusing the data set manual together, the pair found unsatisfactory explanations for how these conflicting variables were defined. They expressed explicit exasperation with the poor quality of the data, as well as the poor documentation in the manual.

In essence, the use of secondary data sets limits the decision space during quantitative analysis. The lack of certain desired variables simplifies selection of variable decisions, as there are fewer variables to select from. This could potentially limit structural or construction decisions as well—i.e., if there were no observed variables to manipulate. Brian's experience highlights this issue, as he was unable to conduct a factor analysis on his observed variables of teacher instructional practices. Although the decision space is therefore artificially limited, the consequence is that researchers are likewise potentially limited in their ability to perform more rigorous analyses that lead to more sound inferences.

CHAPTER 5

METACOGNITION IN STATISTICAL ANALYSIS

The Hastie (2001) model of decision making provided a natural extension to metacognitive processes. With subjective consequences and outcomes driving how people make decisions, metacognition would then consist of evaluating those consequences and seeing how closely the real world outcomes matched the perceived outcomes. If a researcher made a choice that turned out to be a poor decision, the decision could be traced back to the subjective weights given to the consequences and outcomes described in the model. The errors in cognition could then be because the actual outcome did not match an expected outcome or the researcher erroneously favored certain consequences over others. Since subjective consequences and outcomes no longer drive the cognitive process of decision making, however, it stands to reason that other forms of metacognition are necessary to evaluate decisions made by quantitative researchers during data analysis. In this chapter, then, attention is given to the metacognitive process of how researchers evaluate and judge the decisions they have made.

This chapter presents a two-step metacognitive strategy used by quantitative researchers during data analysis. The first step is optional, and occurs when the decision space still allows for multiple specification options, even after the influence of external factors on the decision space. The researchers in this study tended to utilize a cross model comparison, experimenting with multiple models in order to find consistency in results across models. The second step involved developing a story out of the analysis and the results, so that meaning could be derived from the results and decisions made during the course of data analysis could be justified.

Multiple Specification Options

As described in the previous chapter, it is common for researchers to stand on the shoulders of giants when conducting quantitative analyses, and to use the decisions made in previous research to guide them through the decision space. Not all of the territory in the decision space is illuminated, however, and there are many opportunities for researchers to trek through uncharted space. With no research precedence or collective acceptance to help provide justification for particular decisions, how do researchers support and evaluate the decisions that were made? Even after a narrowing of the decision space via the technological restraints and methodological limitations of the researcher, or through the social pressures of a peer group, there may still be multiple available choices to a researcher in analyzing their quantitative data.

Recall from the previous chapter John and Stephen attempting to figure out if it was possible to combine multiple imputation with propensity score matching. Stephen was unsure how to combine these two methodologies, or whether it is even possible or advisable. Ultimately, Stephen decided to "do it every which way" and let the consistency of the results be the criteria by which to judge the results. By that, he meant the use of multiple imputation but also the exploration of alternative model specifications, including listwise deletion. He also intended to try a few variations of propensity score matching and compare the results. He added:

> It's just more work, in that sense that we have to run it ten different ways and just see what the differences are. Once you see how stable those results are, if the results are all over the place, that would be a bad indication that something bad is going on. But if they are all in the same ball park and they all suggest rural is better or rural is lower, whatever, I think that's probably the ideal case. We realize our models are pretty stable in terms of our specifications.

Stephen's approach suggests one metacognitive strategy in quantitative analysis when researchers are unsure of the best option for a particular decision. After he fit multiple models with alternate model specifications, as was suggested, he and John were relieved to find that although parameter estimates and standard errors varied across the different models, the overall findings of significance and ensuring inferences stemming from the analysis did not demonstrably vary across these models. This provided assurance that the conclusion generated from the analysis was reflective of the data itself and not from a poorly specified model.

Eric employed a similar approach for his analysis exploring the effects of governmental assistance programs on high school graduation. One important construct he was measuring was the length of stay in these programs. However, his agency had not previously used data from these programs and thus had no precedent in operationalizing this construct. Eric summed up the dilemma:

> I can use an exit cohort here, or I can use an entrance cohort. But then, what would I do about those that are still in the program? Their data would be censored and you wouldn't have the full layout of how long they stayed. It would underestimate how long everyone was receiving services.

He later added, "I did the analysis for both exit and entrance cohorts," and "I included current consumers in a separate analysis to see if there were different results." Similar to Stephen, one of Eric's strategies to solve the dilemma was to use different model specifications and explore the results across these models.

In Stephen's case, his inclination was to use multiple imputation, as it is considered a more acceptable choice for handling missing observations (Eekhout et al., 2012; Peugh & Enders, 2004). When Stephen compared results across these missing data methodologies, he found few differences in the overall conclusions. As a result, he defaulted to the multiple imputation model since it was more rigorous and allowed for more observations to be included in the analysis. With Eric, however, his multiple analyses yielded differing conclusions. Also, his analysis did not have a methodological approach that is more accepted in academic communities;

using either an entrance or exit cohort would not be considered more or less appropriate than the alternative option. Eric ultimately decided to use an exit cohort for his final analysis. As he explained, "there is less distortion of the data using an exit cohort" and "the analysis with the entrance cohort probably had no findings because it used censored data." When the multiple analyses led to inconsistent results, he used another mechanism to decide the final model. In this instance, data quality was a deciding factor in Eric's final model.

In both of these examples, the analysts used consistency to judge the robustness of differently specified models when a choice in the decision space was unclear. However, this strategy does not imply a particular decision. Stephen and Eric both had other considerations that weighed into their decisions. Whether the alternative specifications produced similar results, as in Stephen's case, or produced conflicting results, as in Eric's analysis, technical aspects of the data and methodologies were still the driving factor in deciding upon the final model specification. Put another way, the strategy of employing multiple model specifications is to be used as a cognitive tool to ascertain the effects of alternate specifications and determine the robustness of findings across statistical models. Other strategies may be necessary for determining the efficacy of decisions made during data analysis.

Storytelling as Metacognition

With the new model of decision making introduced in the previous chapter, consequences of specific actions are omitted and therefore metacognitive process cannot be traced back to subjective outcomes. With the absence of this approach for evaluating decisions, researchers must rely on other means to make sense of their results and to judge the quality of their decisions. After observing researchers conducting data analysis and hearing them talk about their findings, one primary metacognitive process of making sense of results and evaluating the quality of decisions became clear. Specifically, they worked to find the story to tell and in many instances determined whether the results made sense with preconceived theory.

A decision made by Stephen and John further into the analysis helps to illustrate this point. Their analysis required the use of an observed variable to indicate whether a school in their data set would be considered rural or urban. In their data set, there were two such variables. Stephen called one of these variables an “administrative variable.” The variable was provided by the proprietor of the data set; the administrators labeled each school as either rural or urban, and perusal through the data set user guide did not give a clear definition of how either was defined. The other indicator variable of rural or urban in the data set came from the responses of principals who were given the survey. With two variables available to purportedly measure the same construct, the analysts had to make a decision about which of these variables to use in the analysis.

An initial statistical model in the analysis used the administrative variable to represent whether a school was urban or rural. “The problem with the principal variable is missingness—almost 20% missing from the principal,” Stephen said later to justify the decision. Stephen conducted the analysis and fit a statistical model using the administrative variable and then presented the results to John to discuss the findings. The results were unexpected; John expressed frustration with the findings by audibly sighing several times while reviewing the spreadsheet of summarized findings. His frustrations continued as John murmured under his breath, “I just don’t understand why we’re not seeing the results here.” Around this time, Stephen and John both realized that there were a second variable to indicate whether a school was urban or rural—the principal-reported variable.

Stephen produced some cross tabulation tables to compare the responses between these two variables. This table showed that although the administrator and principal agreed in the majority of schools where both variables were available, there was a discrepancy in a large enough portion of schools—almost 2,000 in total—to question whether one or both of these variables was accurate. John and Stephen were at a crossroads. Which of these variables should be used in the analysis? Again, checking the user manual of the data set provided little guidance as to which would be more desirable. After some deliberation, they decided to conduct the analysis for both variables separately and compare the results. The next week, John and Stephen met in John's office to discuss the findings of the second analysis using principal responses. Compared to the previous week, the mood was much more relaxed, as John and Stephen looked over the spreadsheet of results to discuss the findings of the second analysis. The findings were much more in line with their expectations and there were statistically significant results where last week there were null findings.

So at this point in the analysis, a critical decision needed to be made—which of the two variables should be used in the analysis? The pair of researchers had to make a proxy variable decision between two competing measures that supposedly measured the same construct. Earlier in the analysis, one criterion used to assess the appropriate methodology to handle missing data was whether results were consistent with each other. A similar approach was adopted here, and two models were conducted, each using a different variable. Unfortunately, the results of the models led to opposing conclusions. Given the disparate results, coupled with the user manual of the database offering no direction, one would expect anxiety over which variable should be used in the statistical model.

An argument could be made for using the administrative variable to indicate whether a school is urban or rural. There was less missing data compared to the principal version of the variable, and it was not self-reported and thus not susceptible to any personal bias. Moreover, the fact that it came from the administrators of the database, a respected organization with experience collecting and handling a large comparative database, gave it some weight of authority toward its validity. Despite these advantages, John and Stephen decided rather quickly to go ahead with their alternate choice and use the principal-reported variable instead.

How did John and Stephen elect to discard their first model and favor the second model with the principal responses? According to John, "We decided to use this new urban/rural variable and it turns out we're getting the results we've initially expected—the principal reported ones." Without any clear direction as to which of two possibilities was more appropriate for the analysis, the pair of researchers elected for what was consistent with their expectations. The criterion that was used to evaluate their decision was one that supported their hypothesis and was found in prior research. Said another way, the findings matched the story that John and Stephen were attempting to tell.

Other researchers also engaged in storytelling practices to justify decisions made during data analysis. Although they were not as explicit as John and Stephen, the manner in which they talked about their decisions, and especially their results, provides evidence that storytelling is an effective strategy for evaluating decisions made during data analysis. For example, when Brian was exploring teachers' instructional practices and their effect on student achievement, he had to select which countries to include in his analysis. He ultimately ended up using about one third of the countries that were available. Brian explained:

> I need to do this study to try to provide suggestions or recommendations to the policy makers here. So I chose those countries of those education systems that outperform [the]

> U.S. significantly. Also, I chose those countries of educational systems that do not have any significant differences compared with the U.S. Based on those two standards, I chose those 14 countries.

Brian had an incredibly vast number of permutations available but selected a seemingly arbitrary sample out of the available pool to conduct the analysis. His selection was justified by the ability to weave a story out of the findings with the goal of improving policy. Had Brian chosen to select all the countries to include in the analysis, the aggregate findings at the conclusion of his analysis might have been too unclear to develop a story and inform a policy. Selecting a focused sample allowed Brian to concentrate only on relevant countries that were comparable to the United States.

Brian employed a similar storytelling strategy in his selection of variables to examine in the analysis. He did not include all teachers' instructional practices variables in his model, but instead chose to omit one problematic variable:

> That one is about the teachers asking the students to take a written test or quiz. I think that this does not specifically belong to one type of teaching. So in all classrooms the teacher needs to monitor reading tests or quiz. I think this is very ambiguous so I eliminated this one. It does not fit my framework.

Brian decided to omit a variable because it did not fit in with any of the models of teachers' instructional practices that he was analyzing. In other words, the variable did not fit into the story that Brian was attempting to tell and was thus eliminated, even though it partially represented a construct that was of central concern to Brian.

At the end of day, after fitting his statistical model on 14 countries worth of data, Brian was collating responses into a spreadsheet and making sense of the overall results. After spending a few minutes going the significant predictors individually, he offered an aggregate interpretation:

> Initially I have results, in East Asian countries, I found that procedural teaching, memorizing rules, procedures and effects, can be significantly related to students' achievement. This is an interesting result. This result is not consistent in some other countries, but only in Eastern countries, so this is really interesting. For the conceptual teaching, problem solving, NCTM-reform recommended type of teaching, I can see some evidence from this research that can support NCTM type of teaching. Those results can be found in some Western countries. I think that's interesting.

Brian succinctly wove a story from the many significant predictors in the model. This interpretation aligned with the previous theoretical foundations that were the foundation for his current analysis. The ability to interpret results and develop a story out of the aggregate findings is an implicit justification for earlier selection decisions. Eliminating certain countries and teachers' instructional practices variables allowed Brian to find a common thread in the analysis that could be used to inform policy makers. Although he conducted at least 14 statistical models with dozens of predictor variables each, Brian's aggregate conclusion from overall analysis can be summed up rather concisely in just a few sentences. While the story is certainly missing details, the main plot points are definitely already in place.

Variations in Storytelling

The stories used in the analysis of quantitative data may not come solely from the mind of the researcher conducting the analysis, but may be influenced by his or her training, education, and experience. For example, a researcher trained as an economist may look at different elements than a researcher coming from sociology. These two researchers may conduct the same statistical analysis producing identical results, yet they may contextualize the data in different ways and may focus on different aspects of the analysis to tell the story emerging from the results.

Consider Tracy, one of the participants in the current research study who identified as an economist. Tracy was working on an analysis of the effects of a job training program on future salary earnings using a nationally representative database housed by NSF. I observed Tracy preparing her dataset to be ready for analysis, selecting the variables to be included in the model, and writing the code to run the statistical model. Then, it was finally time to read the output of the analysis and make sense of the results. Tracy used her finger to point out the beta coefficients from her multiple regression model on her computer:

> I think this is where the difference between sociology and economics comes from.…Basically, in sociology, they care about a lot of control variables. For us, economists, basically, as long as you add control variables in and the coefficient of what you were interested in doesn't change much, then it's fine. But I guess for sociology, they look at "female are different than males, foreigners are less than citizenship, a lot of people of other race, etc."

What Tracy meant by "control variables" that sociologists are reported to "care about" were other variables in the model—in this particular analysis there were eight total, aside from the job training dummy variable. These eight variables were mainly demographic characteristics, such as gender, ethnicity, and education level. There was also a categorical region variable, split into three separate dummy indicator variables in the regression model. When Tracy said her background as an economist would make her look at just her coefficient of interest, she was essentially choosing to not interpret the other variables in the model.

In these few sentences, Tracy succinctly captured a contrast between her background as an economist and how they may interpret a statistical model with another discipline that may have a different perspective. Researchers from different academic disciplines may look at the same set of statistical results yet see different elements to interpret. That is, different disciplines may focus on different parts of the analysis to build the story; what may be an essential story element for one researcher may not be relevant to the story to another researcher. A covariate in a model can also be an explanatory variable in an identically specified statistical model; it is just a matter of how that variable is conceptualized and woven into a story, which is seemingly influenced by the training and experiences of the researcher. Tracy's observation may be just her own, however. It would be important to seek corroboration from other participants in the research study to determine whether a single analyst is anomalous in her thinking, or whether her suggestion is indicative of an underlying difference between professionals of different academic disciplines.

No other participant in the current research project made a distinction between academic disciplines in the interpretation of analytical results as explicitly as Tracy did. There may, however, be differences in the way that statistical results are absorbed, which may give credibility to Tracy's assertions. Indeed, going back to how educational researchers from other disciplines scanned results from statistical analysis gives some triangulating evidence to Tracy's testimony. Consider Brian exploring effective teacher practices in different countries, or John and Stephen assessing rural and urban differences in student academic achievement. In both instances, these educational researchers scanned the results of their analyses and attempted to weave a story out of the findings, even those that were not the primary focus of their analysis. These researchers examined the significance of other predictor variables in their models and noted any interesting findings that emerged. Covariates were not only used to adjust for bias, but they were also opportunities to explore different components of the overall story.

For example, John and Stephen were finishing up their analysis on rural and urban differences in student achievement. After they discussed the findings for their urban and rural variable across countries, John stated:

> We haven't looked at the other results. There's a lot of interesting results in here that we haven't even talked about yet. Like with indigenous kids. Student work, that's preschool. Look at that preschool is barely significant. Teacher experience…

John's voice started to trail off as he scanned the output for more statistically significant results. A few minutes later, he noticed a significant coefficient for a variable indicating private schools and asked Stephen for help on the interpretation. Stephen noted, "It means there seems to be a private school effect beyond background. So it seems that private schools are doing something that helps in addition to background." Although the primary focus of their analysis was on rural and urban differences in student achievement, John and Stephen were also interested in constructs that affect student achievement as they uncovered elements of their story.

This is contrasted with Eric. Although Eric was at the state education department, his education and background were primarily in economics. His analysis was focused on exploring the effects of receiving governmental assistance on student academic achievement and high school graduation. His analytical model included independent variables such as indicators of being beneficiaries of social assistance programs and the duration that a child's family received these services. The model also included student and family demographic factors, such as race/ethnicity and parental education, and school-level fixed effects. After Eric finished specifying his final model, he began to interpret the results: "The results look like involvement with [government assistance programs] is related to graduation. Duration doesn't seem to have an impact."

After conducting his analysis and interpreting his coefficients of interest, Eric did not revisit other variables in the model, even if they were statistically significant. Unlike John and Stephen, Eric did not seem interested in interpreting covariates, as he was more focused on the primary motivation for his analysis. The contrast between these interactions showcases the differences in storytelling between two qualified researchers. Given Tracy's assertion concerning the differences between sociologists and economists, and the juxtaposition of how

John and Stephen interpreted findings compared with Eric, there is some evidence that there are cross-disciplinary differences in how researchers, even within the field of education, may value different aspects of the analysis, leading to disparate metacognitive storytelling strategies.

CHAPTER 6

DISCUSSION AND IMPLICATIONS

This project was intended as an exploration of the cognitive processes of researchers as they conduct statistical analyses. Borrowing from theoretical developments and research methodologies in cognitive psychology and decision science, the project observed researchers as they analyzed quantitative data to begin to understand the types of decisions they made as well as how they made those particular decisions. The findings led to a revised frameworks that will help develop further the conversation on this topic.

Summary of Findings

Decision Typology

The project was initially framed on the work of Leamer (1978) who proposed a typology of “specification searches” or decisions used by quantitative researchers during the course of data analysis. I used the Leamer typology as a starting point from which to explore the types of decisions made during quantitative analysis. While many of Leamer’s specification searches have relevance today, certain technological and methodological developments in recent decades have made aspects of his typology obsolete and in need of updating. Ultimately, I introduced a revised typology to address some of the limitations of the original Leamer framework.

Rather than forcing Leamer’s (1978) framework onto the current data, I developed a new framework inductively from the observations in the current research project. All observations were analyzed to create a list of the different decisions that were made during the course of data analysis. These different decision points were compiled and categorized to produce a new list of statistical decisions that has plenty of overlap with Leamer’s original framework. Because the current framework was developed inductively, it provides a more appropriate description and a practical list of decisions that were encountered during the course of secondary data analysis.

The new framework built on Leamer's (1978) framework and eliminated specification searches that were overly obtuse. Although there are overlaps with Leamer's framework, the revised framework introduces a typology of two dimensions grounded on the actual decisions made while conducting specific quantitative analyses. The first is the domain of the decision being made. There are five decision domains. *Selection* refers to choosing the appropriate variables and cases in the analysis. A *structure decision* categorizes variables or cases into subcomponents for analysis. *Construction decisions* create groupings of variables, either through aggregating or be creating indices. Decisions related to *model assumptions* are manipulations motivated by meeting assumptions of certain statistical models, such as variables being distributed normally or the lack of outliers in a data set. Finally, decisions related to *generalization* affect the external validity of the analysis—for example, choosing an appropriate proxy variable to represent a construct of interest or appropriate methods for handling missing data to reduce bias of parameter estimates.

These domains are crossed with dimensions specifying whether decisions are related to variables included in a statistical model or the cases being analyzed. The populated cells when decision domains are crossed with variables and cases refer to specific decisions that occur during data analysis. Enveloping these ten decision types is an initial decision regarding the type of model to be used in the analysis. Specific analytical decisions are often dependent on the type of statistical model being used. For example, decisions regarding model assumptions will be driven by the statistical model being used and the assumptions associated with that model.

These decisions are not discrete, and it is likely that a decision in one domain affects a decision in another domain. For example, selecting variables to include when constructing a scale has implications for how to handle missing data as increasing the number of variables may also increase the number of observations with missing data. Some benefits of the revised typology include having a common vocabulary to discuss the different types of decisions that are made during quantitative data analysis, as well as providing a foundation to explore how these decisions are made.

The initial choice typically made by a quantitative researcher is what type of statistical model to use or what type of statistical test to conduct. In many instances, these decisions are made at the commencement of the research process, when the project is being conceptualized. The decision about which statistical model to use is logically derived from the research question being explored and the theoretical foundations supporting the analysis. For example, if a researcher is assessing differences between groups or attempting to assess change over time, this will narrow the decision space of which statistical model to use. The initial model type will also be influenced by the data structures and the type of variables being used in the analysis, particularly when secondary data are used, as they were in the current analysis. These factors also limit the decision space that researchers have in choosing an appropriate statistical model type. Even within these confines, however, there are still plenty of decisions that must be made by a quantitative data analyst, and the initial choice of statistical model will guide decisions deeper into the analysis. For this reason, Figure 6.1 depicts model type as encompassing other decisions being made. Indeed, model choice might be thought of as an umbrella decision that steers subsequent decisions that focus on more granular aspects of the data analysis process.

Figure 6.1 categorizes subsequent decisions along two dimensions. The first dimension, represented in columns, refers to whether the decision pertains to variables or observed cases in an observed data set. This typology is inspired by the display of data in spreadsheets, but it succinctly captures the different elements of data analysis from which decisions are made. Generally, researchers have to decide which variables and cases to include in an analysis, as well as how they will be handled in the model. The second dimension, depicted in each row of Figure 6.1, describes decision categories that pertain to each variable and case. These are the types of decisions that were detected while observing study participants conduct quantitative data analysis.

There were five broad categories of decisions, and each is described in turn below. This discussion includes an explanation of how these categories interacted with the dimension describing variables and cases. Indeed, the analyses will also provide illustrations of these decision types in quantitative research. In the discussion that follows, when there is overlap between Leamer's (1978) typology and the current typology, descriptions of decision types will also refer to Leamer's specification searches.

Model Type

	Variables	Cases
Selection	Covariate Selection	Unit Selection
Structure	Categorization	Subgroups
Construction	Indices/Scales	Grouping/Aggregation
Model Assumptions	Transformations	Outliers
Generalization	Proxy Variables	Missing Data

Figure 6.1. Framework for types of decisions made during quantitative data analysis

Framework for Decision Making

Several scholars in the field of cognitive psychology have conceptualized decision making as selecting between alternate choices and then establishing selection criteria from which to evaluate those choices through a metacognitive process (Chi, 2011; Wang & Ruhe, 2007). Hastie (2001) provided a graphical representation of this process, as previously presented in Figure 2.1. To reiterate, Hastie imagined decisions to be “analogous to a map of forking roads” (p. 656). At the end of each forking road are uncertain events and subjective outcomes associated with each event. The propensity and subjective desirability of each outcome is then evaluated to help decide which direction to take. Although Hastie did not originally develop the concepts being described, the visual representation of the process provides a helpful tool for exploring this cognitive process. This model of decision making will be referred to as the Hastie model from this point forward.

The observations of data analysis and interviews with the study participants identified several external sources that influence how quantitative analysts make decisions in these settings. The next sections describe these sources of influence. Many are social in nature; researchers often rely on the broader community of researchers—either through the direct guidance of colleagues or through previously published research—to help guide their analysis. Other external sources pertain to restrictions of research inputs, either through software limitations or poor quality secondary data sets.

Scholars in the fields of cognitive psychology and decision sciences have conducted numerous studies of how decisions are made in many types of settings and professions. Absent from this line of research, however, is the systematic study of how quantitative researchers make decisions during data analysis. The present research project offered a framework for decision making specifically for quantitative data analysis—one that adapted a general model used for decision making in other fields.

Hastie (2001) provided a visual representation for decision making processes. His model works similarly to a decision tree. In the decision space, multiple options exist for each decision point, each with uncertain outcomes. These outcomes are subjectively weighed for desirability as well as for the probability that they will occur. Decision makers choose the option that has the most satisfactory balance of these subjective outcomes. According to the model, decisions are driven by the perceived outcomes of each branch of the decision tree.

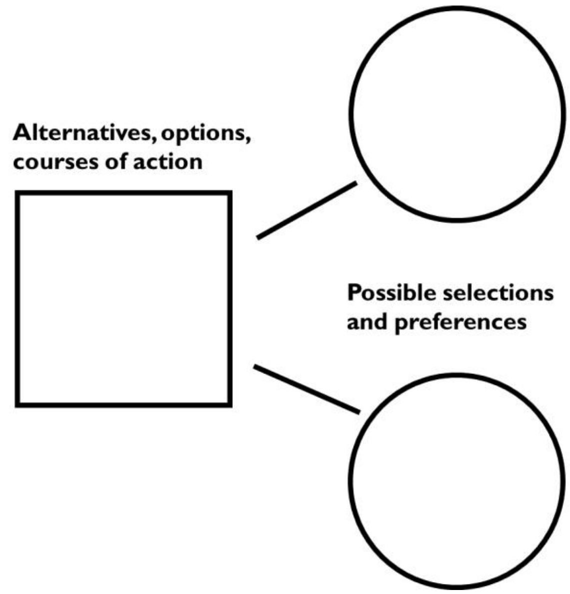
A revised framework for decision making in quantitative data analysis takes into account antecedent factors that influence the branching available at certain decision points. These external factors restrict the decision space by limiting the number of options that are viable. That is, decisions are made not by weighting subjective outcomes, but are rather governed by the direction and influence of external sources guiding the decision maker towards particular options.

The external factors identified in the current study that influence decision making during quantitative data analysis include the role of prior relevant research studies, social factors (including the collective practices of the community of researchers and the peer review process), proficiency capacity dealing with software and methodological limitations, and inadequacies of a secondary data set. These external sources may have differential effects on the decision space, as researchers have different experiences and proficiency levels that may affect their analyses.

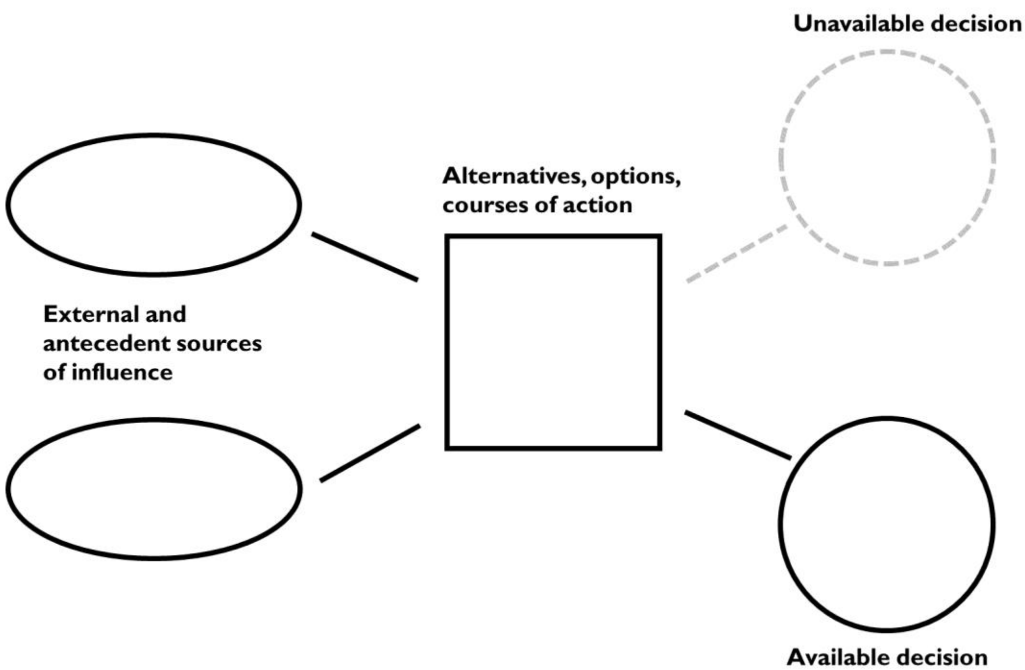
Many of these external sources of influence are social in nature, or at least exist outside the cognition of the decision maker. Although the decision maker has volition, and external sources are not deterministic, external factors play a large role in the decision making process by making some options in the decision space undesirable or even unattainable. Decision making is as much about social psychology as it is cognitive psychology, and limiting studies of decision making only to internal cognitive processes provides an incomplete picture of decision making in quantitative data analysis settings, as this ignores these critical social elements.

Revised Decision Making Framework

The above scenarios demonstrate the limitations of the Hastie (2001) model in describing decision making in the context of quantitative data analysis. In the example of handling missing data, researchers did not subjectively weigh the outcomes and consequences as they chose a specific analytical technique; rather, outside factors influenced their decisions to a greater degree than any perceived cost or benefit. Any serious consideration of the consequences and outcomes of using either listwise deletion or multiple imputation would heavily favor the latter method. However, external factors swayed the researchers towards a certain path in the decision tree, down the road to listwise deletion. They were effectively blind to the option of using multiple imputation. It is evident, then, that the Hastie model is not suitable for describing the cognitive process in quantitative data analysis. The model must be adapted or modified to allow for the presence of external factors that precede a decision and influence how a researcher reacts to a particular decision that must be made during the course of statistical analysis.



Panel A: Decision Space with Possible Selections for a Decision



Panel B: Decision Space with Antecedent Factors Restrict the Decision Space

Figure 6.2. Revised framework for decision making in quantitative data analysis.

The model introduced in the current analysis adapts the Hastie (2001) model to allow for exogenous factors that influence statistical decision making. This adapted model is shown in Figure 6.2. Panel A presents a condensed and modified version of the decision space depicted in Hastie (2001). There is a decision point with possible courses of action. Panel B shows the external antecedent factors that influence the decision point. Rather than having the subjective consequences and outcomes at the end of the decision tree drive how the decision was made, other considerations external to the researcher precede the decision. These antecedent conditions then influence the particular decision point and make certain choices more desirable over alternatives by making them figuratively invisible—as in the case of software and education limitations making multiple imputation a nonviable choice—and through collective acceptance of a particular choice, as in the prevalence of listwise deletion methods. The decision space is effectively narrowed by making certain choices unavailable, and the researcher chooses the decision option that is available.

Similar to the Hastie (2001) model, this adapted model of statistical decision making simplifies the decision space. Decisions are rarely discrete or dichotomous; rather, they may involve choosing between multiple options, and these choices may compound. However, just as the Hastie model can be expanded to include multiple choices with several potential consequences, the adapted model can also include numerous influences that may steer the decision maker towards one of many available choices in the decision space.

The implications of the adapted decision making model are that the researcher has much less agency in making decisions than what is offered by Hastie (2001) or through conventional wisdom. Researchers make analytical decisions less on the merits of perceived consequences or outcomes but are instead swayed through external factors and limitations that restrict the options that are available at any given decision point.

Researchers should be aware of these external factors and of how their analyses might be influenced by these factors. Just because the decision space is restricted through external factors does not necessarily mean that the available options will produce the best outcomes in terms of fitting an appropriate statistical model, as the reliance on listwise deletion attests. It is not improbable that an analysis in the restricted decision space would yield different results and inferences than an analysis with more analytical options available. Different researchers have different decision spaces, just as they have different backgrounds and different social contexts that would affect their available options, even when working on an identical research question with an identical data set. These differences may affect the statistical model and the final results produced.

Metacognition

Metacognition refers to the introspective process whereby individuals assess their own cognitive processes and evaluate the efficacy of decisions made (Veenman et al., 2006). When applied to quantitative data analysis, metacognition represents the process through which researchers judge whether specific statistical model specifications are appropriate for their analyses. The practice of quantitative methodology prescribes several strategies for judging statistical models, such as p-values, test statistics, and model fit diagnostics. It is often the case, however, that these diagnostic tools do not help the researcher decide whether a particular decision was correct.

The current study identified strategies used by researchers to evaluate their analytical decisions. When there are ambiguous options for how to proceed in certain situations, even after restrictions placed on the decision space by external factors, one primary metacognitive strategy is to develop a story out of the research findings. Not only does this provide a method for making sense of the statistical results and contextualizing the analysis, it also offers a justification for some decisions.

When presented with two or more analytical options, one strategy is to conduct the analysis using the available options. This provides a sense of the impact of the decision in terms of the conclusions and inferences derived from the analysis. However, when making the actual choice between these options, the preferred route is the decision that allows the researcher to create a story. Selecting variables and cases for an analysis is only relevant if the selection provides an interpretation that is useful for general knowledge or policy decisions. While storytelling is a useful tool for supporting analytical decisions, there are scenarios where this strategy may produce self-serving biases. When multiple statistical model specifications yield different results, there may be a tendency to select the model specification that most aligns with the preconceived notions and hypotheses of the researcher as it is an easy way to develop a narrative out of the results.

There is also evidence in this study that metacognition varies across researchers, particularly researchers from different fields. Every person has different means of telling stories, and a researcher's background and training presumably play a role in what story elements are desired. Researchers in schools of education tend to look at results holistically and generate a narrative using all elements of the analysis. On the other hand, researchers with backgrounds in economics may look only at specific variables of interest when interpreting results. In this study, covariates were important story elements for educational researchers, but less so for researchers with economics backgrounds.

Implications of the Findings

The current research project initially was framed as an exploration of quantitative researchers through the lens of a cognitive process. As more data were observed, collected, and analyzed, the lens gradually evolved to include social psychology. Cognition of any situation, whether in statistical analysis or otherwise, is influenced by the social context of that particular situation. Ignoring these contexts would provide an incomplete picture of the cognitive processes that occur during quantitative data analysis. The findings of this study have implications for currently practicing quantitative researchers as well as for those teaching the next generation of quantitative researchers.

Implications for Practice

Leamer (1978) developed his initial framework to help guide quantitative analysts by identifying and describing decision points in statistical analysis. Leamer acknowledged the lack of randomized studies in social science research and that inferences based on nonrandomized studies are based, in part, on the specifications made in statistical models (Leamer, 1983). By describing a typology of model specifications, Leamer implicitly illustrates how statistical models can also be misspecified. The results from the current study expand upon Leamer's work by demonstrating the cognitive mechanisms that may lead to model misspecifications.

This project provides support for the claims made by Simmons, Nelson, and Simonsohn (2011) that researchers confirm their research hypotheses—often implicitly—by conducting further analyses when unexpected results occur and stop searching when results align with preconceived notions. Simmons et al. proposed that one mechanism that psychologists may employ is to collect more participants to study when the results are dissonant with their expectations and stop recruitment when these expectations are met. The current research adds an additional mechanism in that researchers may unintentionally confirm their hypotheses by adjusting statistical models until an expected result is achieved.

Quantitative researchers, regardless of their experience level, should consider these influences in their current practice. Any ethical researcher would not intentionally bias their research findings, but bias may creep in in ways that may not be obvious. Remaining cognizant of how the results of an analysis may be challenged by favorable handling of the data can help to reduce self-confirming methodologies.

Another recommendation generated from the research findings involves strategies for justifying analytical decisions when the researcher is unclear about the appropriate model specification. The more experienced researchers in the sample used a cognitive strategy of looking for consistency in findings across multiple model specifications. In Stephen's case, for example, he was unsure of the compounded effects of multiple imputation when combined with propensity score matching. His solution was to fit multiple statistical models, with multiple imputation, listwise deletion, and other methodologies, and then assess the stability of results across these models. Similarly, Shay used multiple classifications to select high achieving students for her sample and Eric used various definitions for operationalizing length of stay in social assistance programs. In these examples, not only was there assurance of findings stemming from the robustness of results, but there were potentially more elements to produce stories from the quantitative results.

When incorporating this strategy into analysis, it is important to recognize the many model specifications and analytical techniques available, whether the researcher is aware of them or not. As methodologies become more advanced and software becomes more technical, it may become easier to remain unaware of analytical strategies that are worth considering. Also, although using specifications from prior, related research may provide ample justification for a particular analytical decision, this precedent may not be the only specification that is viable; alternate model specifications may also be feasible. Thus, researchers should attempt to seek out all reasonable possibilities for alternate model specifications.

The sample of research projects observed in the study varied in the extent to which the projects tested specific theories or were more exploratory in nature. Shay was explicit in stating that her project was exploratory in nature in that she was looking for variables that may be related to high achieving students in mathematics. Other research projects were more theory driven, such as John and Stephen's research study looking at rural and urban differences in student achievement comparatively across countries with different educational practices. The extent to which research studies are theory-driven rather than exploratory plays a role in how results are interpreted and how inferences are made. Research studies that are theory-driven tend to have more story elements in place and the results of these studies are used to fill in the gaps in the story. In contrast, exploratory studies have a lot more flexibility in telling the story from the results and therefore may have fewer opportunities to engage in implicit quantitative analysis practices that confirm existing notions. For example, in John and Stephen's research study, the dyad explore alternate model specifications when the results from an initial model did not support a specific theory that was being tested. Had the pair not come in to the project with preconceived notions and expectations, they might not had questioned the results of the first model and developed a story from that model instead.

However, it may be argued that no analysis, even those that are exploratory in nature, are completely value-free. During the pre-interview, Brian explained that he was looking for relationships between teaching practices and student achievement across countries, a seemingly exploratory analysis. It was only when pressed during the post-interview did Brian explicitly state that he was testing different models of pedagogy. With Shay's analysis, she explains that she is simply exploring factors that are related to high achieving students. However, with her background in teaching mathematics, it would not be inconceivable to believe that she may have prior beliefs as to what factors are related to high performing students. If the notion that value-free analyses are not possible, rather than resisting the impulse to be objective analysts, the community of researchers might be better off embracing their subjectivities as to keep the analyses and the interpretation of results more honest. Then to follow the advice of Simmons, Nelson, & Simonsohn, (2011), make these subjectivities explicit in the reporting of results by informing readers of the decisions made during the course of quantitative data analysis and alternate model specifications that were used but rejected.

It should also be noted, however, that when researchers explore alternate model specifications, the primary objective should be to look for consistency in inferences across these models. Fitting multiple models is not an invitation to select the statistical model that a posteriori confirms the researcher's hypothesis or theory when there are inconsistent results. The multiple models strategy is appropriate for boosting confidence in findings, not for cherry picking results that are most aligned with an expected result. However practical this advice, even experienced researchers are susceptible to picking results that align with their preconceived notions.

These findings would ultimately have implications for public policy as well. As policy makers increasingly rely on evidence based practices and quantitative data to inform public policy, the quality of the data and analyses must be examined to ensure high quality inferences used to inform policy. The findings of this research project suggest that analyses may be susceptible to external sources which may influence the analysis. Furthermore, analysts may implicitly engage in analytical practices that may modify the outcome of such analyses to be in favor of the researcher's preconceived hypotheses. This calls into question the objective nature of quantitative analyses and their ability to inform public policy. Being cognizant of these potential sources of influence may help lead to more objective analyses and ultimately provide more appropriate information to inform matters of social programming.

Even with being mindful of how quantitative data analysis can be influenced, the effects of external sources of influence can never be fully discarded. If it is accepted that bias can still be considerable in quantitative data analyses, then it begs the question of how research can still be an effective tool for making policy. Despite the limitations revealed in this study, quantitative data analysis still can play a significant role in informing public policy decisions by shifting the focus on the type of information that quantitative data analyses can provide. More specifically, quantitative analyses can still be supportive of public policy decisions by moving beyond determining *if* a program or policy is effective and towards illuminating *how* a program or policy is working to benefit its constituents.

Shadish et al. (2002) differentiate between causal descriptions, which state whether a causal relationship exists between an experimental condition and proposed outcomes, and causal explanations, which articulate the mechanisms through which the treatment affects outcomes. With regards to policy analysis, causal explanations extend on causal descriptions by not only determining whether a policy is effective, but detailing why it is effective and under what conditions. Quantitative researchers should seek more causal explanations in their work so that policy makers have more complete information for determining whether to continue or expand a program.

Quantitative analysts can move towards this goal by conducting supplemental analyses and through incorporating theory when interpreting findings. Additional analyses, such as incorporating fidelity of implementation measures into analyses or conducting subgroup analyses can help determine which components of a program are effective and for which specific subpopulations. Social theory often details causal explanations between constructs, explaining how constructs are expected to be in relation to each other. On a more molecular level, specific program theories can explain the processes by which a program is expected to benefit consumers. For example, the introduction of a new math curriculum that is purportedly aimed at improving math outcomes by improving students' engagement with the subject matter. Analyses might incorporate this theory of change by examining whether student engagement is affected by the curriculum and in turn whether increased engagement leads to improved student achievement. Analyses may also look at whether increased engagement and achievement is associated with certain segments of the intended population or with varying levels of fidelity. This would further help describe and explain how the math curriculum is working. By testing the causal mechanism theories and incorporating additional analyses, a more detailed understanding of a program or policy is offered. The increased transparency of the analyses also helps to curb the bias that can influence studies.

Implications for Teaching

The recommendations for currently practicing quantitative methodologists are also applicable to those who are teaching or learning how to apply statistical models. Students are just as vulnerable as seasoned practitioners to letting personal biases influence data analysis, perhaps even more so. As such, they must be aware of the ways in which their results may be affected by various external sources.

One of the core aims of this research project was to examine the essence of the quantitative analyst, or what it means to think like a quantitative analyst. The successful application of quantitative methodology goes beyond the rote memorization of statistical formulas and blind application of these methods to a given data set. It also goes beyond simple dichotomous decisions of statistical significance by examining whether p-values are less than or greater than .05. These are indeed fundamental skills for the quantitative analyst, but they are limited in their ability to provide consequential information to interested parties. Rather, to be a successful quantitative analyst involves giving meaning to the interpretation of statistical results and fitting those results into a proper narrative.

Quantitative methodology instruction should go beyond the mechanisms of particular statistical models and provide a broader landscape of available statistical options. Statistics courses often provide mundane examples of applying a statistical model to a hypothetical data set. This gives the impression that this is the only appropriate statistical model for that data set, when there may be multiple ways to fit a statistical model.

The results of this study shed some insight into what it means to think like a statistician. Quantitative analysts often employ creativity and judgment in their analysis, as evidenced by the analyses included in this research study. This comes from the randomness that naturally occurs in the populations of interest that are under study; certainty and deterministic findings are extremely rare, if not impossible in social science research. Thinking like a statistician involve, in part, accepting this randomness and incorporating it into data analysis and interpreting results. As we have seen in the present analysis, one mechanism for navigating through the randomness is through storytelling mechanisms that help to make sense out of the random nature of social science research. Therefore, one practical recommendation from this research study for students learning statistical methodology is relating theory into their analysis. Not only does this limit the decision space when conducting quantitative data analysis, but it also helps to perceive order from the chaotic randomness that is inherent in social science research.

Brown and Kass (2009) describe statistical thinking as “probabilistic descriptions of variability (p. 107)”. This sentiment is shared by Pfannkuch and Wyld (2000). This means that rather than describing relationships between constructs as deterministic, interpretations from statistical analyses should acknowledge the variability that exists when social phenomenon are under examination. In addition to teaching technical methodologies, statistics courses should incorporate more probabilistic reasoning into the curriculum. This will allow students to better understand that variation exists and hopefully motivate them to pursue options to better understand that variation (Garfield & Ben-Zvi, 2007). Best practices would be to further investigate the variation in statistical results rather than simply interpreting whether a relationship between observed variables is due to chance or not. This provides a more nuanced interpretation of research findings that acknowledges the context of a study and ultimately provides a better understanding of the phenomenon under investigation.

Students of quantitative methodology should be aware of how their future practice may be influenced by the training that they are currently receiving. Different academic disciplines may have varying perspectives on how a particular analysis should be handled. Students should seek a more well-rounded statistics education by actively pursuing these different perspectives. They should also attempt to become proficient in multiple statistical software packages, lest their analysis be restricted by the limitations of certain software packages.

Studies of expertise in other fields have found that one difference between experts and novices in a given field is that experts have a greater ability to pattern match their observations to prior experiences (Chi, Glaser, Rees, 1982). In chess, this might mean recognizing the layout of chess pieces on the board to games that were previously played and making moves accordingly. In statistics, pattern matching might mean developing the ability to decipher diagnostic charts and statistics, such as scatterplots or residual error distributions, and making adjustments in their analysis accordingly. It could also mean matching the pattern of results to previous experiences, in the form of prior research or the relevant literature. This all reflects on the importance of the context of the research study that is sometimes ignored by novice quantitative analysts. Students need to learn about these "soft skills" of data analysis just as much as the mechanisms of statistical models, as the results of quantitative analysis often does not make much sense without being placed within the greater context of the analysis.

Limitations

No research study is perfect, and the current project is no exception. There was much learning that occurred through the implementation of this research project, and this learning reveals the limitations of the study. As previously mentioned in Chapter 3, data collection methods and the sampling criteria evolved during the course of the project. Even with a practice session, think-aloud is not a natural process for most researchers. The think-aloud procedure was intended to allow me to be a proverbial fly on the wall rather than interacting with the data analyst participants. However, most of the researchers had difficulty ignoring me and instead engaged in casual conversation about the analysis. Ultimately this proved to be a productive method for uncovering the cognitive process during data analysis, and by halfway through data collection, I stopped discouraging it. However, this serendipitous discovery meant that earlier participants did not have the opportunity to interact with the observer in the same way.

Another lesson learned from the research project was the variety in the different physical spaces where research analysis happens. Early participants were observed while they were actively analyzing data on their computer. This provided a limited setting to understand the influences that affect how decisions are made. As discovered later in the data collection process, many decisions and discussions take place in social settings—for example, in the office of a professor meeting with his advisee or in a small conference room presenting initial findings to colleagues. By observing some researchers only in front of a computer, I may have limited some opportunities to observe how they made decisions as they analyzed quantitative data.

There were also limitations that were associated with the sample. There was an imbalance in the observed sample and data collection that skewed towards more novice quantitative analysts. The sample included two graduate students, two early career professionals, two more senior level analysts, and a dyad featuring a graduate student and his advisor. Although there were diverse levels of analytical expertise included in the sample, the level and quality of data collection was biased towards the more novice quantitative data analysts; more time was spent on average with the novice quantitative analysts and they provided a richer set of responses during the observations of data analysis and during interviews.

A few factors may explain why novice quantitative analysts provided a greater level of explanation and description than their more advanced counterparts. First, the experts in the sample that were observed for this study had shorter periods of availability. Although the full day was often requested for observation and interviewing, other demands in their workflow often necessitated a delayed starting of their analysis or interruptions in their analysis. Second, the two experts that were observed in the analysis were working on multiple projects concurrent to the projects that were observed in this study whereas the graduate students in the sample were focused primarily on the projects that were observed in this study. This allowed the more novice participants to be more intimate with their project and that familiarity may have led to a richer discussion of that project. Finally, studies of expertise have demonstrated that experts in a given discipline tend to combine problem solving steps, often finding a solution without articulating the discrete steps necessary to arrive at that solution (Chi, Glaser, & Rees, 1982). Although this allows more advanced analysts to solve problems more efficiently, it is problematic when trying to understand the cognitive processes used in solving that problem. A sample that included more expert analysts or with more robust explanations from expert analysts may have yielded different results than those that were found here as experts may engage in different problem solving strategies than their less experienced colleagues (Chi, 2011; Glaser & Chi, 1988).

Every research project is unique in setting and context. The seven data analyses observed for this project should not be considered to be representative of all quantitative social science research. As with any study, replication is necessary to strengthen the conclusions that have been offered. This is particularly the case for the current research project where a limited sample may restrict the range of quantitative data analyses that were being conducted. Although the research findings have helped develop theory about how quantitative researchers think in a general sense, the limited number of observations makes it difficult to generalize beyond those observed in the targeted sample.

Directions for Future Research

This project was intended to be exploratory in nature and was aimed at generating rather than confirming theory. Several claims were generated that merit additional investigation to provide further support for these conclusions. One such conclusion is the role of antecedent external sources that influence how researchers proceed in the decision space. These external sources may interact in ways that are not apparent in the current research project. Future research should explore these relationships and how they influence decisions made by quantitative researchers.

One particularly salient discovery had to do with differences in metacognition across academic disciplines. Storytelling was used as a device to interpret and make sense of statistical output; variations in storytelling elements were discovered across academic disciplines. Future research might attempt to replicate these findings and determine the degree to which academic discipline plays a role in understanding research findings.

The current research project limited the decisions made during the overall research process by observing participants who were working only with secondary data. In this context, many aspects of the decision space were artificially limited by the secondary nature of the data. That is, what was measured, who was measured, and how it was measured were decided by the proprietors of the data. There are several other steps to the research process that precede data analysis and these steps are essentially omitted when only secondary data analysis is considered. Studies must be designed with regards to the setting, the intervention being studied, and the constructs of interest. At each of these steps in the research process, decisions have to be made. Some examples include selecting which participants to collect data from and how to collect data on the constructs of interest. Using secondary data precludes the researcher from having to make most of these decisions.

Yet, similar to decisions that occur during quantitative data analysis, decisions made during research conceptualization and data collection may also vary across researchers and ultimately have an impact of the inferences made at the conclusion of the research. Future research should expand upon the current research by examining how decisions during research design are made and exploring the cognitive processes in planning and developing research studies. It is likely that decisions during research design interact with those during data analysis. For example, choosing the appropriate statistical model is heavily involved with research design. These multi-layered decisions may then compound and greatly affect the conclusions of the research project.

Conclusion

This research project utilized an innovative approach to study a topic that previously had scant research. This is the first step towards a more complete understanding of the scientific process and how it shapes policy. This will also, hopefully, serve as a stepping stone for further introspective scientific inquiry, so that the community of researchers can engage in more credible practice.

Appendix A

Pre-task interview protocol

1. Can you please describe your current position at your university/organization?
    a. *How long have you been at this position?*
    b. *Would you describe yourself as (economist, sociologist, educational researcher, etc?)*
2. What are your primary research interests?
3. Can you describe your previous training and education in quantitative methodology?
    a. *How long has it been since your last course in quantitative methods?*
    b. *How much training did you formally have in quantitative methodology (either through your graduate program or through professional training)?*
    c. *How often do you continue to participate in coursework in quantitative methodology?*
    d. *Have you or do you currently teach courses in quantitative methodology? Please describe*
    e. *To what extent is your knowledge in quantitative methodology self taught?*
    f. *Are there any mentors or influential figures that have shaped your knowledge in quantitative methodology?*
4. How long have you been using quantitative methodology to research issues in education?
5. How did you become involved in analyzing your data set?
    a. *How long have you been interested in comparative education?*
    b. *How long have you been working with this particular data set?*

6. Can you describe the previous analyses that you were involved with in working with this data set or a similar data set?

    a. *What were the goals of your study?*
    b. *What analytical techniques did you use? What techniques did you consider but did not use?*
    c. *What other information did you wish you had access to during the analysis?*

7. Please think back to a point that you got stuck working on a quantitative problem. What solutions did you use to try to solve that problem?

    a. *Did you consult with anybody or consult additional printed resources?*
    b. *What aspects of your training influenced how you approached that problem?*

Appendix B
Post-task interview protocol

1. What was the purpose of the analysis that you conducted with this data set?
2. Can you briefly describe the steps used to create your statistical model?
    a. *If you had to write an abstract describing your analysis, what would you say?*
3. What theory is being tested in your analysis?
4. What variables did you use in your final model?
    a. *What constructs were considered in your analysis? How did you operationalize those constructs using the variables in the data set?*
    b. *What variables were considered but ultimately discard them in your final model?*
    c. *How did you decide that the variable was not a good fit for the model?*
    d. *Did you place any constraints or make assumptions about the variables used in your final statistical model?*
5. To what extent did the analytic process today replicate what you would have done in your own analyses?
    a. *What changes would you have made if you were to have performed the analysis on your own?*
    b. *Was there an outside source that you might have consulted had you performed the analysis on your own? Explain*
6. How did you decide to handle missing data?
    a. *Do you usually use this approach to handle missing data in your other analyses?*
7. (If transformations were used) Why did you transform variables the way that you did?
8. Were there any subgroup analyses being considered?
    a. *What is the population that you generalize your results to?*
9. Why did you include the variables in the model that you did?
    a. *What sources of information did you draw on to determine that these variables were acceptable for inclusion into the final model?*
    b. *Or, what types of information did you draw on to determine that some variables should be dropped from the final model?*
10. Why did you choose the statistical model (i.e., regression, structural equation model, HLM) that you fitted?
    a. *What alternative statistical models might you have considered?*
11. If you had more time to perform your analysis, what other additional approaches or techniques might you have considered?
    a. *What alternate strategies did you consider?*
12. What additional variables or covariates did you wish that you had access to?
13. Briefly, how would you interpret the final model? For example, if you had to present the findings as part of an abstract or executive summary, what would you say?
14. How much confidence would you put in the results of your model? Why?
    a. *How did you judge whether your final statistical model was a good fit to the data? What criteria did you use?*
15. How might the theory you tested be supported or refuted from your analysis?

Appendix C
Coding Framework

Code	Definition
Model Selection	
Alternate Model	An alternate statistical model is conducted on the same data set. This alternate model may take on a number of specifications.
Collapse	Within a single variable or construct, an explanation of how different values are collapsed into a single value
Cost/Benefit	The pros and cons of an alternate model specification are weighed in determining which model might be more desirable
Covariates	A discussion of the appropriate covariates to be used in a model
Index	Explanation of how two or more variables are combined to form a single measure of a construct
Levels	In a multilevel model, a discussion as to which level a predictor would be most appropriate
Model Compare	The results of two or more statistical models are compared to see if there are differences in the interpretation of results
Unit Select	Selecting the appropriate unit of analysis in a statistical test or an explanation of why a particular unit is different from other units and is thus omitted in the analysis
Variable List	Looking through available variables to determine if any might be appropriate for a particular model.
Missing Data	Discussion of the implications of missing data or the degree to which data are missing
Proxy	Using a similar variable as a substitute for another variable that may have been more desirable
Transformation	When transformations are used on a variable to better fit a statistical model
Outliers	Screening for outlying observations or decisions regarding how to handle outliers
Conceptual	
Construct	A conceptual discussion of a construct and its definition and how it might be operationalized
Future Research	Generating ideas for future research or explorations based on the current analysis
Hypothesis	A statement of the predicted results of a statistical analysis

Code	Definition
Operational Definition	A broad construct of interest is defined in a manner that is observable and measurable
Explanation	
Alternate Explanation	Explaining the results of a statistical model using a different interpretation
Assumption	Explicit assumption about the manner in which a variable is defined or how it is used in a statistical model
Contemplate	Musings regarding an alternate approach to a statistical model or how the results might be interpreted
Definition	Discussion about how a particular variable is defined and measured in a secondary data set
Explanation	The researcher explains the technical process of what the statistical model or adjustment is purportedly doing.
Interpret	Interpreting the results of one or more statistical analyses in everyday language and without statistical terminology
Justification	An explanation as to why a particular decision was made in a statistical model.
Results	
Holistic	The overall pattern of results is examined and interpreted; Results are examined in aggregate rather than individually.
Negative Finding	The results of an analysis are the opposite of what is hypothesized or theorized by the researcher
Null Finding	The results of an analysis reveal that there are no statistically significant differences or a coefficient is not significant; the null hypothesis is not rejected
Positive Finding	The results of an analysis supports the theory or hypothesis of the researcher
Unexpected	The results of a particular analysis elicits a reaction of surprise from the researcher
Subgroup	Opposite of holistic, considering the results of each subgroup individually rather than aggregated
Superlative	An expression of astonishment or excitement over the result of a statistical analysis
Previous Experience	
Prior Analysis	The researcher refers to a prior analysis personally conducted to help guide the current analysis
Prior Knowledge	Refers to the researcher using prior substantive knowledge of a particular subject to inform the analysis or interpret results
Prior Research	The researcher refers to an analysis conducted by somebody else to inform the current analysis

Code	Definition
Social	
Colleague	Seeking or considering the help of a colleague to help understand a statistical model or its results
Discipline	Discussions about how researchers in specific academic disciplines generally think about a certain topic or approach a statistical model
Journal	Consideration of journal reviewers or journal editors in the consideration of statistical modeling
Software	A discussion regarding the limitation of a particular software being used or being unfamiliar with particular software and how that may affect analyses
Secondary Data	
Data Quality	Discussions around the limitations of particular variables, either in their definition or measurement, in a secondary data set
Measurement	Consideration of how a particular qualitative variable is defined or measured
Unknown	When the reasoning behind a lack of a particular variable or how the variable is defined in a secondary data set is unknown

References

AbouZahr, C., Adjei, S., & Kanchanachitra, C. (2007). From data to policy: Good practices and cautionary tales. *The Lancet, 369,* 1039-1046.

Alkin, M., Christie, C., & Rose, M. (2006). *Communicating evaluation.* In I. Shaw, J. Green, & Mark, M. (Eds.), *Sage handbook of evaluation policies, programs, and practices.* (pp.384-403) London: Sage.

Atran, S. (1998). Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences, 21,* 247-609.

Baker, J. J. (1997). Effects of a generative instructional design strategy on learning mathematics and on attitudes towards achievement. Dissertation Abstracts International, 58(7), 2573A.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29,* 37-65.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28,* 51-75.

Bartholomew, D. J. (1995). What is statistics?. *Journal of the Royal Statistical Society, Series A (Statistics in Society), 58,* 1-20.

Bickel, R. (1989). Post-high school opportunities and high school completion rates in an Appalachian state: A near-replication of Florida research. *Youth & Society, 21,* 61-84.

Bergee, M. J. (2005). An exploratory comparison of novice, intermediate, and expert orchestral conductors. *International Journal of Music Education*, *23*, 23-36.

Bickel, R. & Papagiannis, G. (1988). Post-high school prospects and district-level dropout rates. *Youth & Society, 20,* 123-147.

Bickman L., & Reich S. M. (2009). Randomized controlled trials: A gold standard with feet of clay? In Donaldson S. I., Christie C. A., Mark M. M. (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 51-77). Thousand Oaks, CA: Sage Publications.

Brown, E. N., & Kass, R. E. (2009). What is statistics? *The American Statistician, 63,* 105-110.

Borko, H., & Livingston, C. (1989). Cognition and improvisation: Differences in mathematics instruction by expert and novice teachers. *American Educational Research Journal, 26,* 473-498.

Boruch, R. (2005). Preface: Evaluation for evidence-based policy: Place randomized trials in education, criminology, welfare, and health. *The ANNALS of the American Academy of Political and Social Science, 2005,* 6-18.

Campbell, R. J., Kyriakides, L., Muijs, D., & Robinson, W. (2004). *Assessing teacher effectiveness: Developing a differentiated mode.* London: Routledge Farmer.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55-81.

Chi, M. T. H. (2011). Theoretical perspectives , methodological approaches, and trends in the study of expertise. In Y. Li, & G. Kaiser (Eds.), *Expertise in Mathematics Instruction*. New York: Springer.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.) *Advances in the Psychology of Human Intelligence.* Hillsdale, NJ: Erlbaum.

Cogan, L. S., Schmidt, W. H., & Wiley, D. E. (2001). Who takes math and in which track? Using TIMSS to characterize U.S. students' eighth-grade mathematics learning opportunities. *Educational Evaluation and Policy Analysis, 23,* 323-341.

Cook, L., Cook, B. G., Landrum, T. J., & Tankersley, M. (2008). Examining the role of group experimental research in establishing evidence-based practices. *Intervention in School and Clinic, 44*, 76-82.

Cook, T. D., Scriven, M., Coryn, C. L. S., & Evergreen, S. D. H. (2010). Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. *American Journal of Evaluation, 31*, 105-117.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs.* San Francisco, CA: Jossey-Bass.

Cronbach L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Eekhout, I., de Boer, M. R., Twisk, J. W. R., de Vet, H. C. S., & Heymans, M. W. (2012). Missing data: A systematic review of how they are reported and handled. *Epidemiology, 23,* 729-732.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis: verbal reports as data*. Cambridge, MA: MIT Press.

Fife, B. L. (1992). Communicating the results of evaluation studies. *Science Communication, 14,* 234-256.

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistics Review, 75,* 372-396.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis.* London: Chapman and Hall.

Gersten, R., & Hitchcock, J. (2009). What is credible evidence in education? The role of the What Works Clearinghouse in informing the process. In Donaldson S. I., Christie C. A., Mark M. M. (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 78-95). Thousand Oaks, CA: Sage Publications.

Glaser, R. & Chi, M. T. H. (1988). Overview. In M. T. H. Chi, R. Glaser, & M. J. Farr. (Eds.), *The Nature of Expertise* (pp. 287-310). Hillsdale, NJ: Erlbaum.

Glaser, B. G. & Strauss, A. L. (1967). *The discovery of Grounded Theory: Strategies for qualitative research.* Chicago: Aldine Publishing Company.

Hastie, R. (2001). Problems for judgment and decision making. *Annual Review of Psychology, 52*, 653-683.

Hedges, L. V., Stodolsky, S. S., Mathison, S., & Flores, P. V. (1986). Transition Mathematics Field Study. Chicago: University of Chicago.

Henry, G.T. (October 2000). Using evaluation findings for policy: A realist perspective. Paper presented at the European Evaluation Society Conference. Lausanne, Switzerland.

House, E. R. (2008). Blowback: Consequences of evaluation for evaluation. *American Journal of Evaluation, 29,* 416-426.

Jones, B. K., & Mitchell, N. (1990). Communicating evaluation findings: The use of a chart essay. *Educational Evaluation and Policy Analysis, 12,* 449-462.

Kritzer, H. M. (1996). The data puzzle: The nature of interpretation in quantitative research. *American Journal of Political Science, 40,* 1-32.

Latour, B., & Woolgar, S. (1979). *Laboratory Life: The Social Construction of Scientific Facts*. London: Sage Publications.

Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review, 73,* 31-43.

Levelt, W. J. M., Drenth, P., & Noort, E. (Eds.). (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel.* Commissioned by the Tilburg University, University of Amsterdam and the University of Groningen.

Little, O., Goe, L., & Bell, C. (2009). A practical guide to evaluating teacher effectiveness. National Comprehensive Center for Teacher Quality.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* New York: Wiley.

Lundgren-Laine, H., & Salantera, S. (2009). Think-Aloud technique and protocol analysis in clinical decision-making research. *Qualitative Health Research, 20,* 565-575.

Lyle, J. (2003). Simulated recall: A report on its use in naturalistic research. *British Educational Research Journal, 29,* 861-878.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29,* 67-101.

McDonough, P. (1997). *Choosing colleges: How social class and schools structure opportunity*. Albany, NY: State University of New York Press.

Nielsen J, Clemmensen T, & Yssing C. (2002). Getting access to what goes on in people's heads? reflections on the think-aloud technique. Proceedings of the second Nordic conference on Human-computer interaction, October 2002.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, 115, Stat. 1425 (2002).

Peugh, J. L. & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74,* 525-556.

Pigott, T.D. (2001). A review for methods for missing data. *Educational Research and Evaluation, 7,* 353-383.

Pfannkuch, M., & Wild, C. J. (2000). Statistical thinking and statistical practice: Themes gleaned from professional statisticians. *Statistical Science, 15,* 132-152.

Raudenbush, S.W., and Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage Publications, Inc.

Rockoff, J. E. & Speroni, C. (2011). Subjective and objective evaluations of teacher effectiveness: Evidence from New York City. *Labour Economics, 18,* 687-696.

Schoenfeld, A. H. (2006). What doesn't work: The challenge and failure of the What Works Clearinghouse to conduct meaningful reviews of studies of mathematics curricula. *Educational Researcher, 35,* 13-21.

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science, 26,* 113-125.

Scott, C. (2005). Measuring up to the measurement problem: The role of statistics in evidence-based policy making. PARIS 21 partnership in statistics for development in the 21st century. London School of Econmics.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Shavelson, R. J. & Towne, L. (Eds.). (2002). *Scientific research in education*. National Research Council. Committee on Scientific Principles for Educational Research. Washington, DC: National Academy Press.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011) False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science, 22,* 1359-1366.

Stough, L. M., & Palmer, D. J. (2003). Special thinking in special settings: A qualitative study of expert special educators. *The Journal of Special Education, 36,* 206-222.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research.* Newbury Park, CA: Sage Publications.

Swanson, H. L., O'Connor, J. E., & Cooney, J. B. (1990). An information processing analysis of expert and novice teachers' problem solving. *American Education Research Journal, 27,* 533-556.

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., Ariet, M., Fisher, T., & Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29,* 11-36.

Thompson, D. R., Senk, S. L., Witonsky, D., Usiskin, Z., & Kaeley, G. (2005). An evaluation of the second edition of UCSMP Transition Mathematics. Chicago: University of Chicago School Mathematics Project.

Torres, R. T., Preskill, H. S., Piontek, M. E. (1997). Communicating and reporting: Practices and concerns of internal and external evaluators. *American Journal of Evaluation, 18,* 105-125.

Tourmen, C. (2009). Evaluators' decision making: The relationship between theory, practice, and experience. *American Journal of Evalaution, 30,* 7-30.

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think-aloud method: A practical guide to modeling cognitive processes.* San Diego, CA: Academic Press Ltd.

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning, 1,* 3-14.

Watanabe, T. (2011, March 28). "Value-added" teacher evaluations: L.A. Unified tackles a tough formula. *The Los Angeles Times.* Retrieved from http://www.latimes.com/news/local/la-me-adv-value-add-20110328,0,3903343.story.

Weiss, C. H. (1988). Evaluation for decisions: Is anybody there? Does anybody care? *Evaluation Practice, 9,* 5-19.

Weiss, C. H. (1999). The interface between evaluation and public policy. *Evaluation, 5,* 468-486.

Westerman, D. A. (1991). Expert and novice teacher decision making. *Journal of Teacher Education, 42,* 292-305.

Williams, R. W., Roberts, G. W., Irvine, F. E., & Hastings, R. P. (2010). Exploring decision making in intellectual disability nursing practice: A qualitative study. *Journal of Intellectual Disabilities, 14,* 197-220.

Wolfe, E. W., Kao, C., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication, 15,* 465-492.