Determining the Importance of Design Elements for Secondary Analysis

A Thesis

Presented in Partial Fulfillment of the Requirements for the

Degree of Master of Science

with a

Major in Statistical Science

in the

College of Graduate Studies

University of Idaho

by

Mitchell Davies

Major Professor: Michelle Wiest, Ph.D.

Committee Members: J.D. Wulfhorst, Ph.D.; Jeremy Kenyon, M.S.

Department Administrator: Chris Williams, Ph.D.

July 2015

Authorization to Submit Thesis

This thesis of Mitchell Davies, submitted for the degree of Master of Science with a Major in Statistical Sciences and titled "Determining the Importance of Design Elements for Secondary Analysis," has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: _____ Date: _____

Michelle Wiest, Ph.D.

Committee Members:_____ Date: _____

J.D Wulfhorst, Ph.D.

_____Date: _____

Jeremy Kenyon, MS.

Department

Administrator : _____Date: _____

Chris Williams, Ph.D.

Abstract

Data collection is an important step in the scientific process with much of a researcher's time and funding spent on collecting the necessary data to understand their field of study. Unfortunately, after the data collection is complete, data often is not made available to other researchers. In order to promote reproducible research as well as re-use of existing data, there is growing trend of data sharing. Documentation is one of the important aspects of data sharing which is used to aid in the understanding and reuse of shared data. The importance of what type of documentation to include for secondary analysis is vital especially documentation about the design of the study. This thesis analyzes what design elements are the most important to include in this documentation for making data more reproducible and reusable for secondary analysis.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Data collection is an important step in the scientific process with much of a researcher's time and funding spent on collecting the necessary data to understand their field of study. One issue that frequently occurs when collecting data is that often this data are not made available to other researchers. As a countermeasure for this issue, a growing trend has been emerging called data sharing. Data sharing is when researchers make their data publicly accessible enabling other researchers to check, reproduce and reuse the data that they had collected through secondary analysis. Thus, researchers benefit by reducing the time and costs of data collection, and cut down the amount of redundant data. Supporting this trend many government funding agencies including the United States Geological Survey (USGS), National Science Foundation (NSF), and others, now require their researchers to share the data they collect (Holden 2013). In collaboration with the USGS Climate Science Centers, our research team was tasked with evaluating the best practices needed in order to help share social science data which was starting to be collected by the climate science centers. During this evaluation, the importance of including documentation with shared data was frequently brought up.

Documentation, or sometimes called metadata, is the information about the data. Examples of documentation include: what sampling method that was used, what population was studied, who collected the data, etc. Documentation provides important contextual information about the data that helps researchers use the shared data in their analyses. Unfortunately, issues often arise with the amount of documentation that is provided with shared data. In interviews conducted with data management experts we found that the most common problem in data management is that researchers do not provide enough documentation with the data they share. Interestingly, when asked what requirements or recommendations they provide to researchers only a few actually made any requirements for documentation.

Thus, we see a conundrum resulting in minimal, or no documentation being provided with the data. Without good documentation about the data, secondary analysis would be very hard if not impossible, especially without documentation about the study

design. These design elements are all used by researchers in the data collection process and affect the type of analyses that can be done using the data. To help understand what design elements are needed to be shared this study investigated what design elements are the most important for making data more reproducible and reusable for secondary analysis.

This work was funded by the USGS NCCWSC.

Chapter 2: Literature Review

We conducted a brief review of the literature pertinent to research design. When selecting articles for review we organized the articles into three categories: 1) research design, 2) secondary analysis, and 3) data curation. Each of these categories was selected to help better understand different aspects of the research process with research design representing the data collection process, secondary analysis representing the analysis process, and data curation representing the archiving process. When reviewing the literature, the main focus was to find out what type of design elements are referred to for each category and how they are used. Rather than focusing on all fields of study, we focused our research on articles about the social science field.

Research Design

Research Design is the foundation of any good study and affects every aspect of the research. Thus, understanding what design elements that are considered important is vital. This importance is mentioned by Daly (2003) when she wrote:

"Overall it is very important to realise that the methods and techniques one chooses are part of a broader package. It is quite widely accepted that methodology involves a set of standards which should be aspired to. Less widely acknowledged is the fact that assumptions and values underlie all methodologies as well as a particular view of how we are to understand the social world. We need to be as conscious of the assumptions and conditions attaching to our methodology, as we are in applying and using them" (Daly, 2003).

In reference to this, failure to understand the study's design can lead to problems in analysis. Thus, we first need to look at the data collection process to understand what design elements need to be focused on.

Research design used by researchers often vary from field to field and researcher to researcher. "Research design is *not* the step-by-step procedures one goes through in carrying out a piece of research" (Miller, 2003). This makes it difficult to determine what

type of design elements are important. When discussing research design most articles focus on a few aspects, and often only mention a handful of design elements that were used by that researcher specifically. This is especially apparent when a study's focus is based on quantitative data or qualitative data exclusively. Not only does using quantitative data vs qualitative data differ in the methods collected but also with the type of design elements that are considered important. For example, quantitative data focuses on information about the structured data set (e.g. variable name, and variable labels), where qualitative data documentation focuses on interview protocol (e.g. processing the interviews transcription, and coding results). Even with these differences, there were some common elements that were mentioned in the literature. Our review revealed five categories of design elements used for social science: general study elements, collection method elements, implementation method elements, input method elements, and analysis method elements.

General study elements are design elements that refer to basic information about the data. These design elements are normally sets at the beginning of the study (Miller, 2002). These elements are typically used to help understand what a researchers study is about and helps explain other aspects of the study design. Design elements that would be classified as general study elements are research questions, sub questions, the population being sampled, underlying theory, etc.

Collection method elements are design elements that describe how the data are collected, and what collection formats are being used (Bechhofer, 2000). Design elements in this category includes the process for collecting the data (e.g., surveys, interviews, experimentation, observation, or using a mixed methods approach). It is important to know the medium in which the data was collected. Depending on the type of collection method, certain design elements in other categories will be considered important. For instance, if you were using a survey, the question type used has an effect on what type of analysis that can be performed on the data.

Implementation method elements are design elements that describe the process used to reduce bias (Bloom, 2005). Design elements in this category include sampling

methods, sample frame, sample size, etc. Expressing the value of randomization, Ioannidis (2014) wrote, “For randomised trials, allocation concealment, blinding, and method of randomisation might modify effect estimates, especially for subjective outcomes” (Ioannidis, 2014). Because of this possible modification on the effect estimate it is important to have this information to make sure that researchers reusing the data will be able replicate the analysis.

Input method elements are design elements that describe how a dataset is cleaned and organized by the researcher. This is normally seen in a codebook and includes variable names, variable labels, measurements used, etc. (Litwin, 1995). The design elements used in this section can be different depending on the collection method. For example, interviews are typically presented in a coded transcription so including a transcription and code list is important when sharing this type data. Since most researchers are going to see and use a structured version of the data, it is important to include everything that these researchers might need to help them use the data set.

Analysis method elements are design elements that describe how the data was analyzed. Design elements in this category include any type of analysis and code used by the researcher. Certain types of secondary analysis use this category more than other. For example, if a researcher is reanalyzing data to check another researchers results, the analysis method used becomes important information to include about data. Overall, there is a lot of different design elements that can have effect on a researcher’s analysis. This means it is important to include these design elements to make sure that the data are understood for the use in secondary analysis

Secondary Analysis

Secondary analysis is analysis of data where the researcher who is analyzing the data are typically not the primary researcher who collected the data. Secondary analysis allows researchers to complement primary research and analysis with their own perspective on the topic (Hakim, 1982). Some examples of secondary analysis include combining other

researchers' data with the researcher's primary dataset to increase the quantity of data giving researchers a larger, more representative sample of the population. Other researchers' data may also be combined to provide a longitudinal perspective giving researchers a historical picture of the data. This can be especially useful for social science considering a society often changes its perspectives, and ideals over time. Being able to see the changes that society goes through helps to give researchers a better perspective on the evolution of ideas and culture. Thus, secondary analysis allows researchers to make inference on a population from a unique perspective with little to no data collection.

The challenge with secondary analysis is that researchers using the data may not properly understand the data since they are not the primary researcher. The problem is that "[d]ata on deposit are only as useful as the information about them makes them" (Tanenbaum, 1980, p. 34). Because of this, it is crucial for a researcher to have a good understanding about how the data was collected. Louise Corti, in reference to qualitative research wrote, "The most significant issue currently being debated in the consideration of secondary analysis is of that data and original context. The basic argument lies with the belief that qualitative data cannot be used sensibly without the accumulated background knowledge and tacit understanding that the original investigator had acquired in 'being there'" (Corti, 2008). Thus, without a good understanding of the data it is challenging, and perhaps inappropriate to analyze the data.

To fully understand secondary analysis it is imperative that we look at the different types of secondary analyses. There are many different types of secondary analyses. For example, a researcher might be trying to reproduce the results of another study, or a researcher might be trying to increase their sample size to decease their uncertainty about the population estimate. When reviewing the articles on secondary analysis there are two different groups of secondary analyses; 1) those that are using a single dataset, and 2) those using multiple data sets. Single dataset analysis includes analysis to reproduce results of another study, and analyses to answer your own hypothesis about the sampled population. The design elements needed for single dataset analysis are simple. For example, when reproducing someone's results the design elements are the type of analysis, and any

information that would help with this analysis like sample weights, or sampling methods. Other design elements that seem common with single dataset analysis, are general study elements and input method elements (e.g. variable labels, variable units, population sampled, etc). These elements are used to make sure researchers fully understand the data they are using.

Multiple dataset analysis includes combining datasets like combining a dataset with the Census, combining datasets to increase sample size, or meta-analysis. Meta-analysis occurs when one combines analysis results called effect sizes (Wolf, 1986). Looking at multiple dataset analysis, we see the combination of design elements considered important are more complex than single dataset analysis. The situation is that with multiple dataset analysis researchers are trying to combine datasets by using variables that are the same or similar to each other. This means researchers are trying to pick datasets that are very similar to each other. Similarities could be making sure that similar populations are being sampled, checking the type of sampling methods that were applied, making sure that variables that we are combining are using the same units, or checking combined studies or variables will answer the hypothesis researchers are asking.

If two datasets are similar, then they are called homogeneous datasets. Homogeneity of the studies examines the similarities of two studies or datasets (Higgins, 2011). This does not mean that the studies have to have the exactly same design elements; they just need to be *similar enough* studies. The more homogeneous the datasets are, the easier the datasets are to combine. Also, having homogeneous datasets help to reduce the bias and error that can occur during analysis. After reviewing the difference types of secondary analyses we found two common uses of design elements in secondary analysis: 1) selecting the data to be used, and 2) reducing bias in your analysis.

When it comes to selecting the data for secondary analysis, design elements are used to help make sure that the data will be answer the hypothesis the researcher has in mind. Most of these design elements are from are general study elements, collection method elements and input method elements. Some examples include population

information, location collected, type of data collected and many elements from the codebook which includes variables labels, and units used.

With any study there is bound to be bias or errors that will occur. There is always uncertainty when performing analysis, and this can increase for secondary analysis when we don't have a good understanding of the data we are using. When trying to reduce bias and uncertainty in an analysis, specific categories of design elements seem to be used. This categories are collection method elements, implementation method elements, and a few input method elements. Some example include sampling method, weights used, and imputation methods. Now that we have seen what documentation are considered important for both research design, and secondary analysis, we reviewed what type of documentation are used in data curation.

Data Curation

Data curation is managing data so they are reusable and available to other researchers. In reference to the goal of data curation Macdonald wrote, "these records need to be maintained so that the data and information remains usable for as long as required or wished –possibly extending over decades or more –so that we can discover, locate, retrieve and re-run the data with confidence and at appropriate cost"(Macdonald, 2006, p.116). Thus, we see having as high quality data are important for data curation. When reviewing articles about data curation, an interesting problem about the quality of data was noticed.

Articles about data curation seem to be divided about the true quality of curated data. The problem is that the quality of curated data are considered as both an advantage and a disadvantage depending on the articles you read. As already discussed, understanding the data are important to help us to complete secondary analysis, however the articles reported mixed results as to whether curated data provided this information. On the one hand, many articles referred to archived data as being a high quality data product with all the information you could ever need is available to you (Greenstein, 2006). This idea comes from the fact that archivists, when getting data are collecting as much documentation as

possible and optimistic researchers will provide this data. When reviewing articles that consider the data a high quality data product not much information about specific design elements were included. These articles are imply that if you need any information it would be available. This idea comes from the fact that data curators use templates called metadata standards. Metadata standards are schemas for documentation, and many articles would say that if a metadata standard was applied, then it would have the right amount of information.

Other articles refer to the data provided by archives as weak and not well documented (Schwartz, 2013). The problem resided in the fact that not all data has the same research methods and not all fields of science use the same methods. Referring to this Cragin (2010) said "these research communities tend to be heterogeneous in the methods and data types applied, without uniform or widely applied data standards, and are not currently well supported by disciplinary repository services. It is anticipated that these scientists will require access to a wider range of curation services to support deposit data into shared repositories." (Cragin, 2010, p. 4025). This leads to the question of what design elements need to be included in order to make sure researchers are making available a high quality data product. Like with secondary analysis when looking at data curation, we identified two main uses of design elements. Design elements were either used for preservation of the data or to help with reuse of the data.

Documentation is need to help to make sure that the data will be useful to researchers for years to come. Documentation that helps with preservation includes format types, codebook information, etc. There is more information that could be included about preservation documentation, but the focus of this paper is more about secondary analysis and reuse so we will be focusing more on that aspect.

When looking at design elements used for reuse of data Corti wrote, "Comprehensive and accurate documentation is essential for informed and accurate use of the data; thus, data should be accompanied by file and contextual documentation that describes how the data were created (including sampling and fieldwork practices), prepared

for analysis (e.g., transcribed, digitized), and subsequently collated. The content of each data file, such as an interview, should be recorded, and the record should include information about who was being interviewed, when, where, and so on." (Corti, 2008). We see that many of the design elements that are considered important for reuse are also important for research design and secondary analysis. These design elements include the research question, sampling methods, sample frame, collection method, etc. The only difference that we saw was that they were not only focusing on one field of science and so even though there were many pieces of information that were the same there were other documentation that were also included (e.g. documentation on data security and locations, and getting access to the data or dealing with sensitive data). In the end, we see that for each of the three groups there are similarities and differences. The similarities for the three groups can help to understand that the most important design elements are universally research question, selection criteria, collection method used, sampling method, sampling weight, variable labels, and measurements used.

Chapter 3: Evaluating Elements of Design

To further narrow down what design elements are needed for secondary analysis we used a mixed methods of analyses. This mixed methods approach used three types of data collection and analyses: 1) interview with data management experts, 2) survey of social science researchers about data sharing and 3) simulations to show the effects of ignoring research elements in secondary analysis. The purpose of these three analyses was to understand what researcher's and expert's attitudes are toward data sharing, and to illustrate the pitfalls of ignoring design elements in secondary analysis. This chapter will summarize the collection methods used and review the results of each of these analyses.

Interviewing Data Management Experts

We interviewed 21 experts in the data management and data archiving fields to give us a base understanding of data sharing. We selected the interviewees from a sample frame created from employees of top data management and archiving services (e.g. data manager, archive managers, documentation specialist, etc). Using recordings of these interviews we transcribed the interviews into written form. These transcriptions were reviewed and coded to identify main themes that were mentioned by the interviewees in response to the interview questions. From these codes, we were able to create a final code list based on multiple transcripts. After recoding the transcripts, we calculated frequencies of the themes.

In these interviews, our focus was to understand common problems seen, and any common requirements and procedures used by data management experts. As mentioned before, when asked what the most common problem that is seen in data management and curation, 87.5% said lack of documentation. When referring to lack of documentation, topics mentioned include data having incomplete metadata, disorganized or incorrect codebook, or lacking certain fields of documentation. Interestingly, when experts were asked about what policies and procedures are required or recommend before researchers make their data public only 42.9% had requirements for metadata, and a mere 14.3% recommended a codebook.

Other interesting results we found were when we asked about use of different types of data processing performed on social science data. Specifically, we looked at how researchers deal with sensitive data. We found many process these type of data very differently, with 38.1% said deleting sensitive data, 33.3% said recoding sensitive data, 19.0% said hiding sensitive variables, and 14.3% said anonymize sensitive data. Using this information as a guide, we created a survey to administer to social science researchers.

Surveying Social Science Researchers

We created our survey based on themes found from the interviews. These themes were used to inform response items and questions that were asked in the survey. We sampled social scientists who have worked with biophysical data or vice versa. We sent out 94 surveys and received 49 completed surveys. Our sample frame started with researchers from the USGS climate science centers who had worked on social science projects for the USGS. Using a snowball sampling method we sent a link through email we provided subjects we a web survey to complete. To aid in the response rate of the survey we attempted three follow up calls with the sampled researcher who had not completed the survey. Thanks to these follow up calls, we were able to get a strong response rate of 52.13%.

To ensure that we were surveying the right demographic, we first asked the participants do they consider their primary field of study to be social science, biophysical science, or other. We found that 72.3% of participants said their primary field was ‘social science’, 10.6% said ‘biophysical science’, and 14.9% said ‘other’ (Figure 1). We therefore had a good representation of social scientists that participated in the survey.

Our next focus was on trying to understand participant’s attitude and experience when it comes to data sharing using two groups of questions. First, we asked about their attitude toward data sharing which includes survey questions such as “how important is sharing their data with others?” and “how important is documentation of shared data?” These were setup as 5 item likert scale questions ranging from ‘very important’ to ‘not important’. When asking about the importance of data sharing in their discipline the majority of researchers responded that data sharing was ‘very important’ with 36.6% or

'moderately important' with 22.0% (Table 1 and Figure 2). Similar results were found when asking about the importance of sharing (Table 2 and Figure 3), and when looking at how important documentation is to sharing data (Table 3 and Figure 4). From all this a simple pattern emerges showing that when talking about attitudes about sharing data, people consider data sharing very to moderately important to researchers.

Secondly, we asked researchers about their experiences with data sharing which include survey questions such as "how often do you share your data with others?" and "how often do you use data that others have shared?". These questions were designed as 5 item likert scale ranging from 'never' to 'always'. We found that the majority of researchers responded 'often' with 29.3%, 'sometimes' with 36.6% or 'rarely' with 19.5% (Table 4 and Figure 5). Similarly when asked about how often do they used data shared by others the majority of responses were 'often' with 24.5% and 'sometimes' with 39.0% but not 'rarely' which only got 2.4% (Table 5 and Figure 6). This data seems to focus on middle of the scale results showing most researchers experience with data sharing as only being 'sometimes'.

After asking about their attitude and experience with data sharing, we asked about what type of design elements they have used. Survey questions for this included "what type of social science data do you collect and/or used?", "What type of collection methods have they used in the past?", "What type of sampling methods have you used in the past?", and "What design elements do you consider important to include when sharing data?". All of these questions were multiple answer questions where the participants could select all that apply based on a list of different common options.

We asked researchers what type of data they collected. The options included 'audio', 'imagery', 'video', 'geospatial', 'quantitative', 'qualitative', and 'none' and respondents were allowed to chose multiple types. The most common responses of data type were 'qualitative' 87.2%, 'quantitative' with 80.9%, and geospatial with 61.7%, (Table 6 and Figure 7). When asked about type of collection methods researchers have used in the past we provided the options: 'interview', 'survey', 'focus group', 'panel dialog', 'ethnographic', 'observations', 'content analysis', 'multimedia', 'other', and 'none'. We

found that top results were ‘interviews’ with 85.1%, and ‘surveys’ with 83%. Other notable results but not as common were ‘observations’ with 57.4%, 'content analysis’ with 55.5%, ‘focus groups’ with 55.3%, and ‘ethnographic’ with 44.7% (Table 7 and Figure 8).

When asked about what sampling methods have researchers used in the past, we provided the options: ‘simple random’, ‘stratified’, ‘weighted stratified’, ‘systematic’, ‘cluster’, ‘quota (non-random)’, ‘convenience’, ‘other’, and ‘none’, The most common sampling methods used were said ‘simple random’ with 66%, ‘stratified’ with 57.4%, ‘convenience’ with 57.4%, and ‘systematic’ with 40.4%. (Table 8 and Figure 9). Notably these are not very high frequencies, just the most common. Lastly when asked what design elements researchers considered important we provided the options: ‘subject collection criteria’, ‘sampling frame’, ‘location collected’, ‘mode of collection’ (phone, email, …), ‘method of collection’ (survey, interview,…), ‘sampling method’, ‘sampling weights’, and ‘other’. We found that 87.2% said ‘method of collection’, 80.9% said ‘location collected’, 78.7% said ‘subject selection criteria’, 72.3% said ‘sampling method’, and 63.8% said ‘sampling frame’ (Table 9 and Figure 12).

Next, we looked at different types of data processing that are normally done with social science data. Specifically we looked at how researchers deal with missing data, and sensitive data. To do this we asked the questions “how often do they have missing data in their datasets”, and “how sensitive are the data they collect”. Each question was setup with a 5 item likert scale with the question on missing data ranging from ‘never’ to ‘every time’, and the question about sensitive data ranging from ‘not at all’ to ‘extremely’. When looking at how often researchers see missing data we found the most common responses were ‘almost every time’ with 28.9%, and ‘occasionally’ with 42.2% (Table 10 and Figure 10). When looking at how sensitive researcher’s data are we found that similarly the most common answer were ‘moderately’ with 22.5%, ‘neutral’ with 30%, with other notable results on the lower end of the spectrum where 20% said ‘slightly’, and 17.5% said ‘not at all’ (Table 11 and Figure 11).

Following this we asked what type of methods they use to deal with missing data

and sensitive data. The questions that are used were “what is the most common method for dealing with missing data”, and “what are the most common methods for dealing with sensitive data”. For the question looking at the most common method for dealing with missing data we asked the subject to pick one out of the four options which were ‘statistical imputation’, ‘deletion of records’, ‘special coding’ or ‘no action’. We found the results were mostly even, with 29.8% said ‘special coding’, 21.3% said ‘statistical imputation’, 21.3% said ‘deletion of records’, and 14.9% said ‘no action’ (Table 12 and Figure 13). When looking at the most common methods for dealing with sensitive data we set it up to be a multiple answer question with options including ‘top coding’ (suppressing extreme values), ‘deleting sensitive information’, ‘recoding/anonymizing’, ‘collapsing categories’, ‘statistical disclosure control’, and ‘other’. We found that ‘recoding/anonymizing’ with 51.1%, and ‘deleting sensitive information’ with 40.4% were the most common methods used (Table 13 and Figure 14).

After looking at the individual results of each question we went into the analysis of the questions together. We used cross tabulation of the questions to see if there is a pattern with the type of study they are using and what design elements they consider important or methods that they use. The purpose of this is to see if any of the frequencies previous stated were due to some other factor that might have not been taken into account. So these combinations are split up into two groups the first looking at whether the type of data affected the sampling method used, the method of dealing with missing data, the method of dealing with sensitive data or the design elements that are considered important. Since there are a lot of items being combined, cross tabulation table are included with the plots for each table. For looking at data type compared to these different sampling methods refer to Table 14 and Figure 15, for data type compared to the method of dealing with missing data refer to Table 15 and Figure 16, for data type compared to how sensitive is researcher’s data refer to Table 16 and Figure 17, and lastly for data type compared to what design elements are considered important refer to Table 17 and Figure 18.. When reviewing these results it is important to note that some of the groups only had a few responses. If more response were available for these groups a pattern might have

emerged. There was no discernable pattern between the variables in our dataset.

The second group of cross tabulations similarly looks at the method of data collection compared to the sampling method used, method for dealing with missing data, method for dealing with sensitive data, and what design elements are considered important. This was done to check if the collection method affects the methods and design elements used. Again since there are a lot of items being combined, cross tabulation table are included with the plots for each table. For collection method compared to sampling method used refer to Table 18 and Figure 19, for collection method compared to method of dealing with missing data refer to Table 19 and Figure 20, for collection method compared to how sensitive is researcher's data refer to Table 20 and Figure 21, and lastly for collection method compared to what design elements are considered important refer to Table 21 and Figure 22. Similar to the type of data no discernable pattern was seen.

Simulating Secondary Analysis

Following this survey, we used simulation to see what type of effect leaving out documentation had on secondary analysis. We specifically wanted to simulate ignoring two different types of design elements sampling methods and missing data processing. We first looked at what would happens if you tried combining datasets without knowing what sampling method was used. At the start of this simulation, we created two different populations to test. Both populations dataset had 3 variables: city, gender and annual wage. The city and gender variables were selected the same way for both populations with the city being set first and gender being randomly set with a 50% probability. As for annual wage, the two populations used different methods to simulate the variable. The first method was to simulate a uniform population, we used one gamma distribution, with α being 80,000 and β being 2, to simulate annual wage. The second method was to simulate a more real population, each city and each gender had their own gamma distribution to simulate annual wage. These gamma distributions had α's ranging from 140,000 to 45,000 and all the β's being 2. After this, we simulated five studies on the population, using randomized sampling methods; simple random sampling, systematic sampling, stratified

sampling or cluster sampling. The sample size was also randomized from 50 to 200 for each study. These samples were then combined using three methods; 1) the melting pot method, 2) the mean of the means method, and 3) weighted means method.

The melting pot method is where you combine data as though they did not come from separate samples. When calculating the confidence interval using this method we assume that simple random sampling was used for each study (Yen, 2002). The mean of the means method combines the datasets by the means of the samples. This method starts by:

1. Calculating the means $\bar{x}_i$ of each sample separately using the appropriate method.
2. Calculate the mean $\bar{\bar{x}}$ using the formula $\bar{\bar{x}} = \frac{\sum_{i=1}^{m} \bar{x}_i}{m}$, where m is the number of dataset combined.
3. Calculate the variance $var(\overline{\overline{x}})$ using the formula $var(\bar{\bar{x}}) = \frac{\sum_{i=1}^{m} (\bar{x}_i - \bar{\bar{x}})^2}{m-1}$.
4. Calculate the confidence interval using the formula $\bar{\bar{x}} \pm t^* \sqrt{var(\bar{\bar{x}})}$, where $t^*$is the critical value from the Student's t distribution with $m - 1$ degrees of freedom (Yen, 2002).

Finally, the weighted means method using the inverse of the standard error as a weight to calculate the mean and variance. This method starts by:

1. Calculate the means $\bar{x}_i$ and variance $s_i^2$of each dataset separately using appropriate methods.
2. Set raw weights to be $w_i = 1/(s_i^2/n_i)$, where n_i is the sample size of each individual datasets.
3. Calculate mean using formula $\bar{\bar{x}} = \sum_{i=1}^{m} w_i * \bar{x}_i / \sum_{i=1}^{m} w_i$, where m is the number of dataset combined.
4. Calculate variance using formula $var(\bar{\bar{x}}) = \sum_{i=1}^{m} (w_i / \sum_{i=1}^{m} w_i)^2 * s_i^2$.
5. Calculate confidence interval $\bar{\bar{x}} \pm t^* \sqrt{var(\bar{\bar{x}})}$, where $t^*$is the critical value from the Student's t distribution with $m - 1$ degrees of freedom (Yen, 2002).

Using these methods, we calculated three 95% confidence intervals and compared them to

the true population mean. This process was repeated 10,000 times giving us the frequency of how often the confidence intervals contain the true population. We also created boxplots of the difference between the true mean and the calculated means to check for possible bias.

For each population, we replicated this entire process three times (Table 22). For the uniform population we found that using the melting pot method 92.9, 96.0, 97.3%of 95% confidence intervals contained the true mean with the average margin of errors being 9.22, 9.23, and 9.22 giving us very small confidence intervals. Also when calculating the mean there was no bias (figures 24-26). For the mean of the means method all three trials showed 100% of the 95% confidence interval contained the true mean with the average margin of errors being 234369.2, 230906.4, and 231291.6 giving us very larger confidence intervals. Also when calculating the mean there was no bias (figures 27-29). Finally, for the weighted means method 95.74, 95.35, 96.09% of the 95% confidence intervals contained the true mean with the average margin of errors being 1867.8, 1764.6, and 1758.8 giving us good sized confidence intervals. Also when calculating the mean there was no bias (Figures 30-32). From this we see that for completely uniform population, all methods seem very good to calculating the population estimate even though the size of the confidence intervals were very different.

Next, for the more real population in which city and gender influence income, we found more varied performance (Table 23). For the melting pot method 16.4, 20.4, 20.3% of 95% confidence intervals contained the true mean with the average margin of errors being 821.9, 821.5, and 820.5. Also When calculating the mean there was some bias with the bias being around 1865.8, 1849.5, and 1895.6 (figures 33-35). For the mean of the means method, again we found all three trails showed 100% of the 95% confidence interval contained the true mean with the average margin of errors being 419693.1, 421507.9, and 420634.2. Also when calculating the mean there was no bias (figures 36-38). Finally, for the weighted means method 72.99, 73.22, and 72.0% of the 95% confidence intervals contained the true mean with the average margin of errors being 11185.7, 11183.38, and 11129.0. Also when calculating the mean there was bias with the bias being around 4062.6, 4200.8,

4243.7 (Figures 39-41). This helps us to see the true effects of ignoring the sampling method. The melting pot method which ignored the sampling method only really worked with the ideal population. When the population was more realistic we saw that the melting pot method had large under coverage and some bias. The mean of the means method didn't have any under coverage or bias but the margin of error was so large that it does not tell us anything about the true population. Lastly, the weighted means method had some under coverage with some bias but had more reasonable confidence interval sizes. From this we see the importance of choosing the right method and the effects of not taking into account the sampling method.

Lastly, we performed a simulation to show the under coverage that comes from dealing with missing data. We picked mean imputation method to use and wanted to show the under coverage and bias that is created using this method. For this simulation, we sampled using simple random sampling five times and then randomly deleted from these samples a number of observations. We then used mean imputation and replaced the missing data with the mean values and calculated confidence intervals from this repaired dataset. We saw that 88.4% of combined studies contained the true mean in their confidence interval showing the potential under coverage that comes from using certain methods of dealing with missing data. To show this in more detail we created a box plot of difference between the true mean and combined means see if there was any bias in the calculation of the means and we saw that there was not (Figure 23).

Chapter 4: Conclusion

Looking at literature review and analyses we can see the importance of documentation in data sharing. The majority of researchers consider data sharing and documentation important even though they don't always share their own. We also see this importance when looking at the simulation work done. We illustrated the potential biases and under coverage that can come about if you do not collect enough information about the data.

What is stopping researchers from sharing their data? Is it too much effort to submit documentation with the data? The policy of "collect as much documentation as possible" is not working. We see from the literature there are so many different design elements to consider, it is not practical to ask for every piece of information. Therefore, the gap this study addresses is figuring out what design elements are truly important for documentation.

At first glance, this question seems like it would be easy to answer, but we determined through this study that there are many different aspects of the research design that we need to be taken into account. Through the literature review and analyses, many different design elements were identified, making it difficult to pinpoint the most important elements. This idea is fortified by the results from the survey with many researchers saying that they use both quantitative and qualitative collection methods. Even when asked about design elements they considered important most of the choices had similar high frequencies with a few exceptions. Interestingly, those few exceptions can be considered important at time. Sampling weights was one of the exceptions in the survey that was not considered very important. However, as we saw from the simulation and what we know about stratified sampling, the sampling weights used can be very important to your analysis.

From this study we concluded that data curation is like publishing a paper. There are many general rules to follow, however, some of the rules are more due to preference or circumstance. For the general rule pertaining to design elements we need to be able to answer the questions "what is the researcher trying to study?", and "How was the data collected?" Design elements like the research question, location collected, selection criteria,

collection method used, sampling method, sampling frame, and some design elements from the codebook are important. These elements help to set the general rule of design elements to include as all studies are going to have used some form of these elements in their study design.

The design elements that are not as clear are those that are used some of the time in studies or only used with specific collection methods. For example, sampling weights, interview procedures, code list, and processing methods. For these, it becomes important for researchers and data managers to communicate and figure out if these design elements are needed. Data is an important scientific contribution and with any type of good contribution it takes time and effort to create. With good communication between researchers and data managers the problem of lack of documentation can be overcome and archives will be able to provide researchers with higher quality data for secondary analysis

Appendix A: Figure

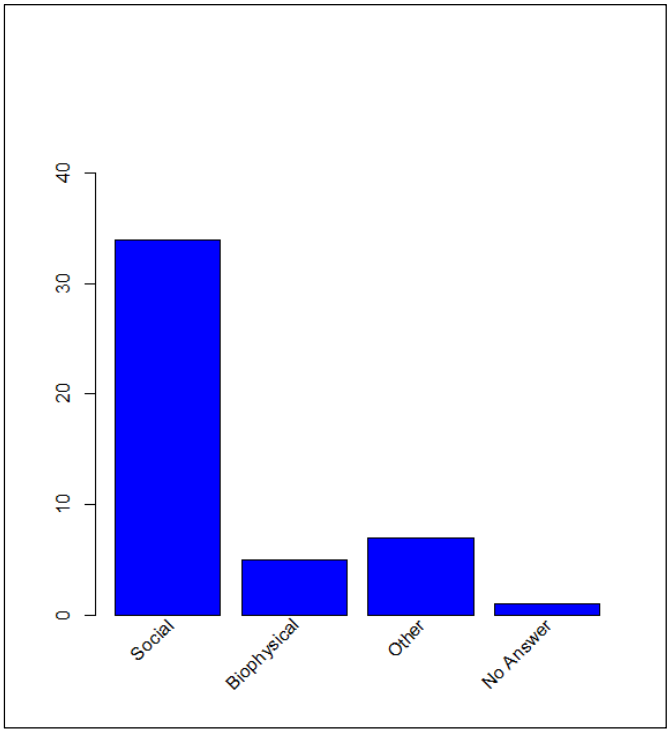

Figure 1: Field of Study Bar plot

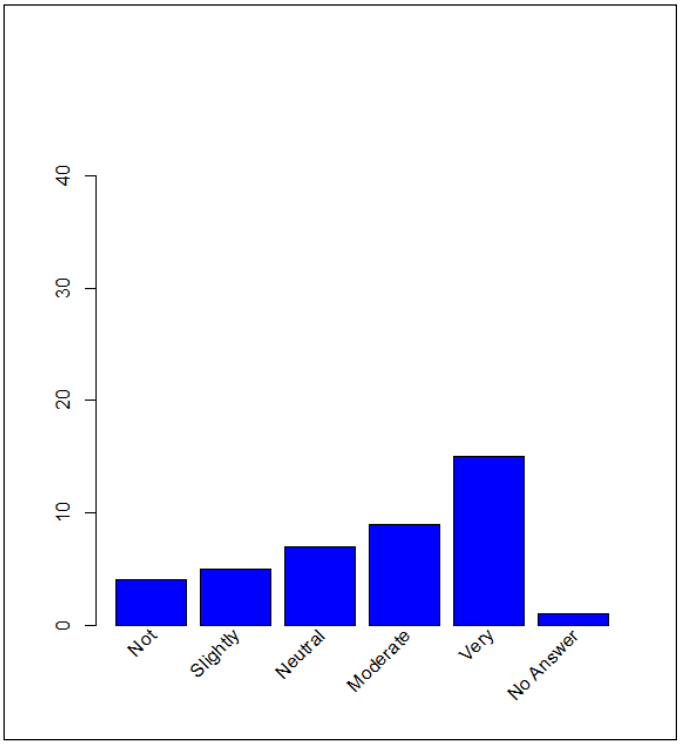

Figure 2: Importance of Data Sharing in Discipline Bar plot

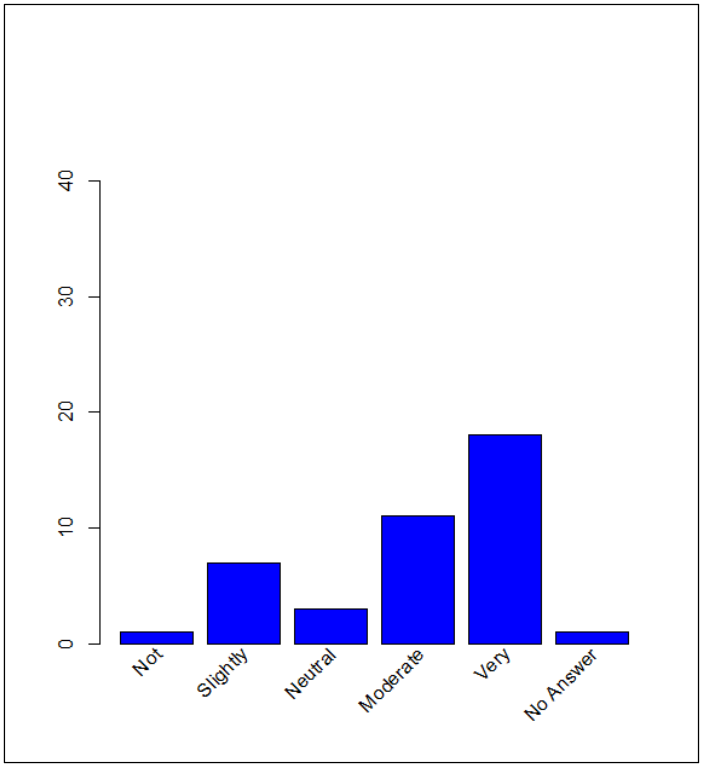

Figure 3: Importance of Data Sharing Bar plot

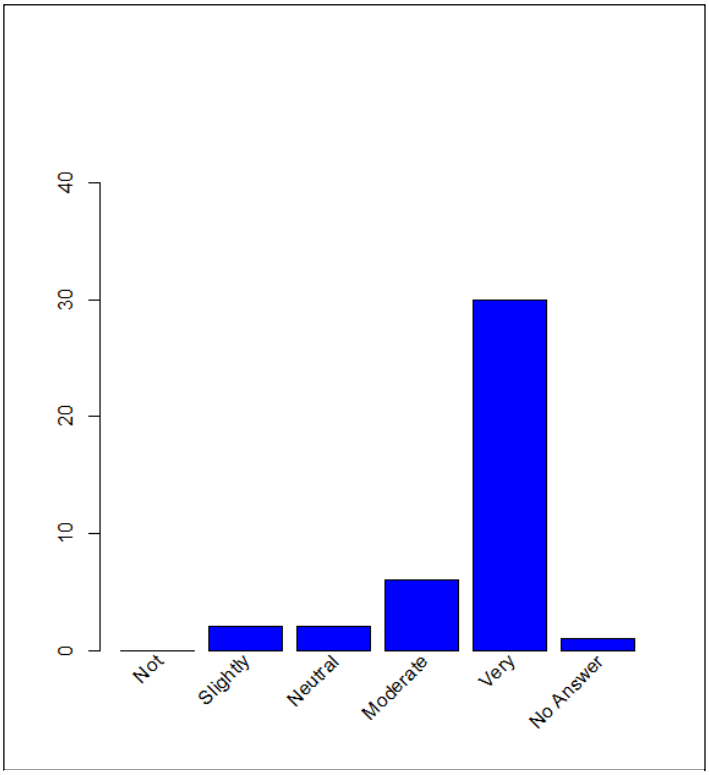

Figure 4: Importance of Documentation Bar plot

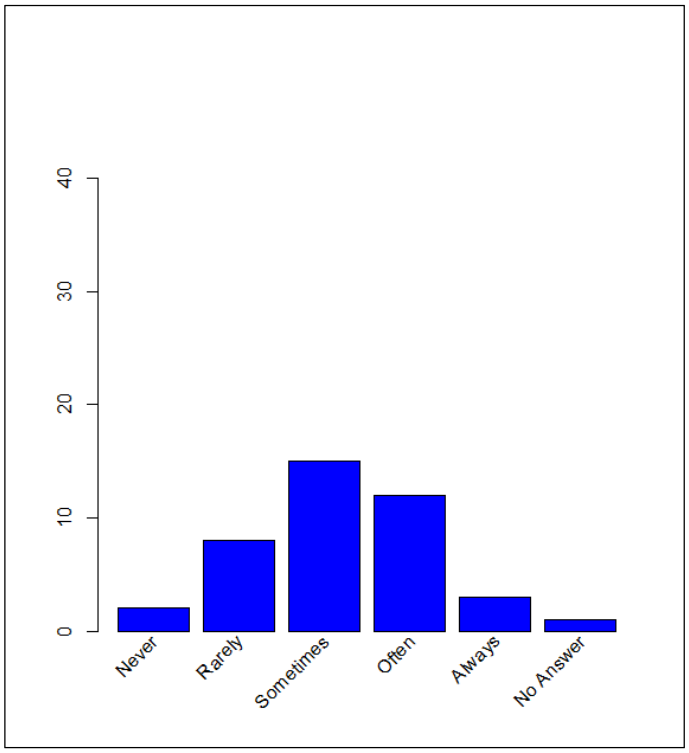

Figure 5: Experience with Sharing Data Bar plot

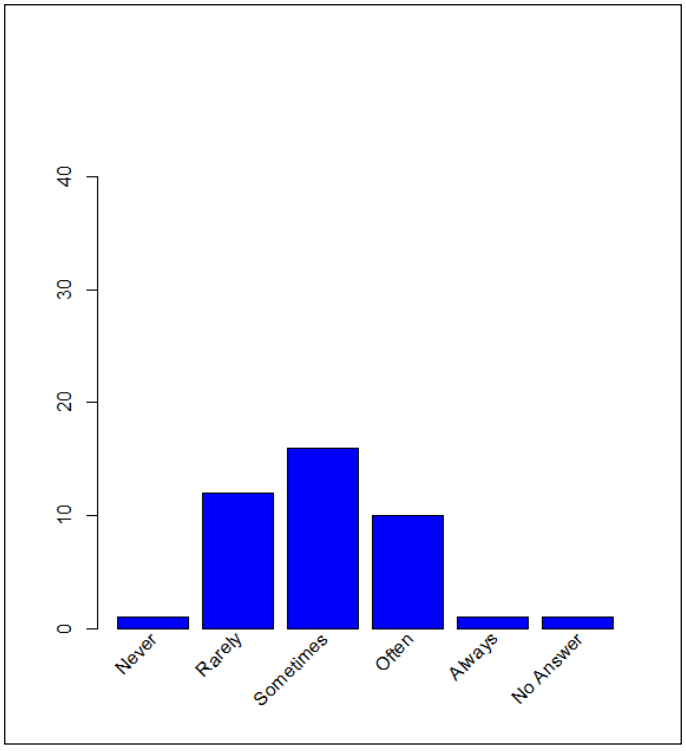

Figure 6: Experience with using Shared Data Bar plot

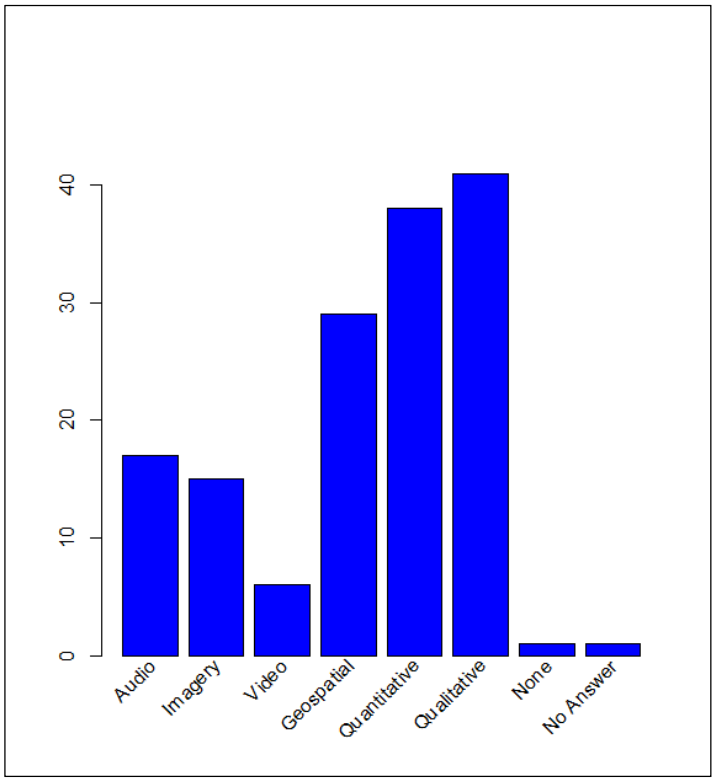

Figure 7: Type of Data Bar plot

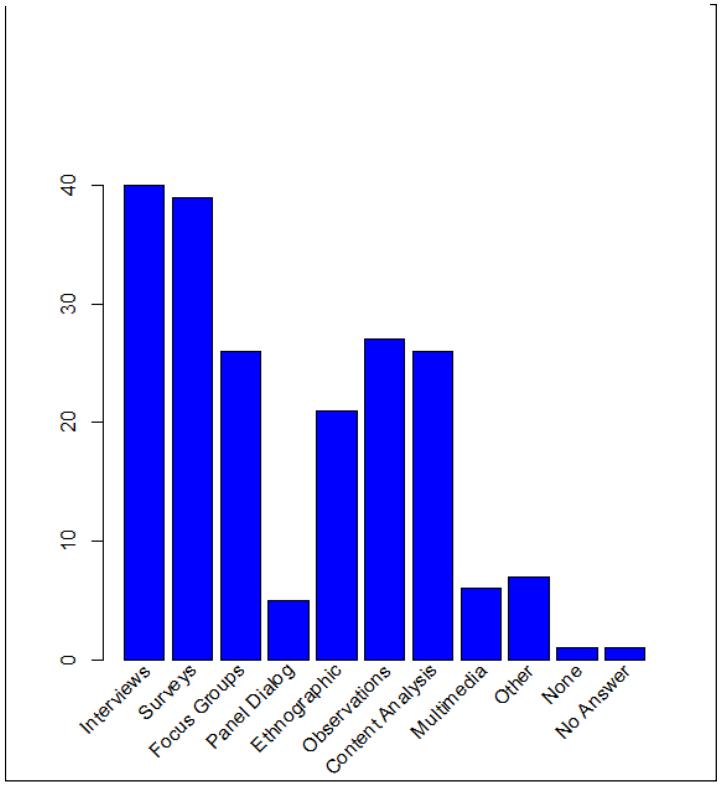

Figure 8: Collection Method Bar plot

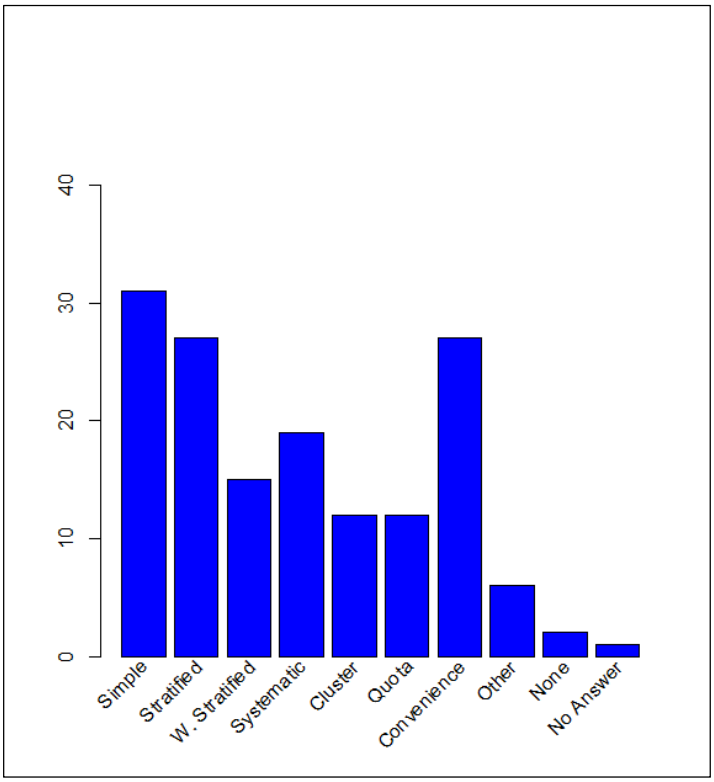

Figure 9: Sampling Method Bar plot

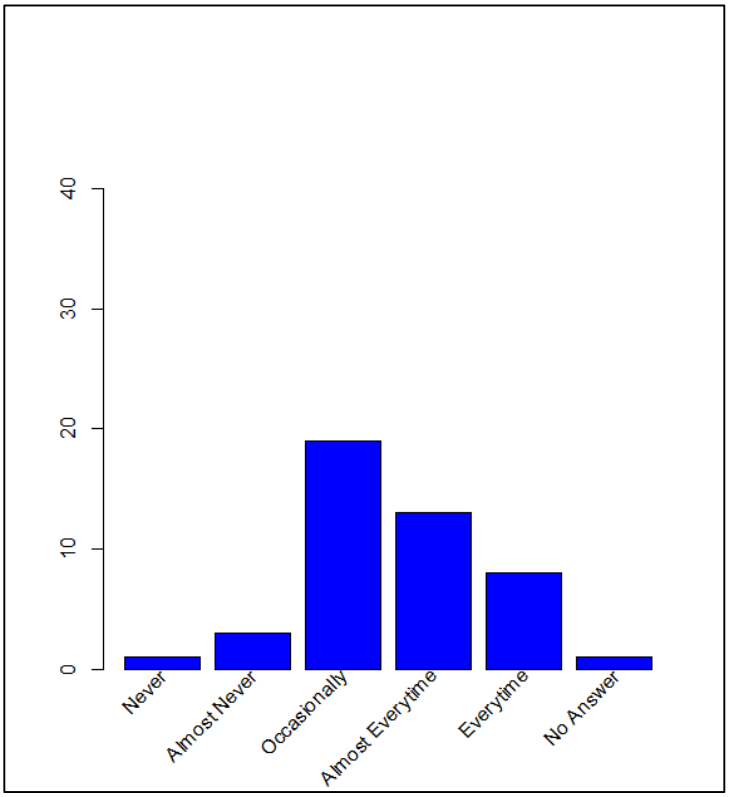

Figure 10: Frequency of Missing Data Bar plot

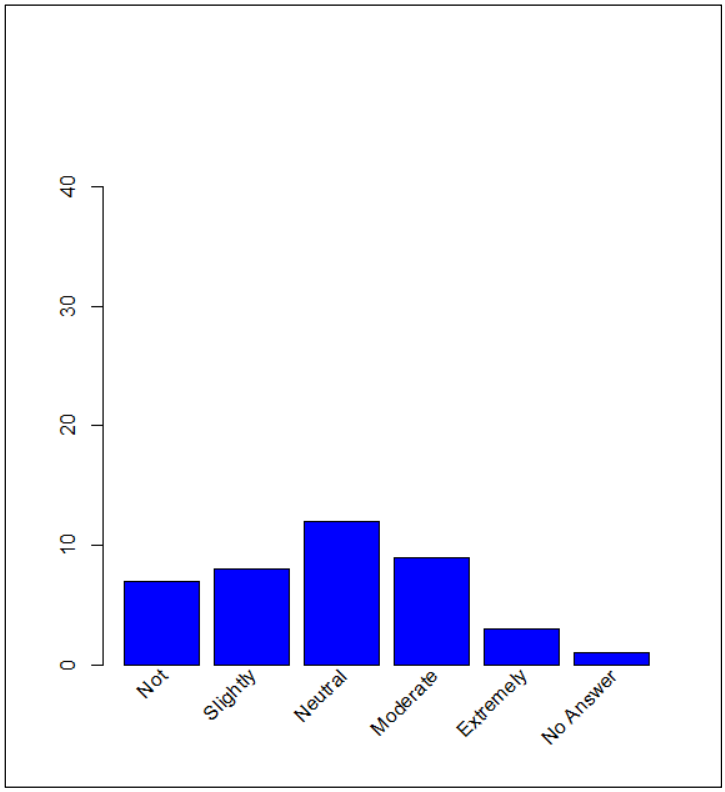

Figure 11: Sensitivity of Data Bar plot

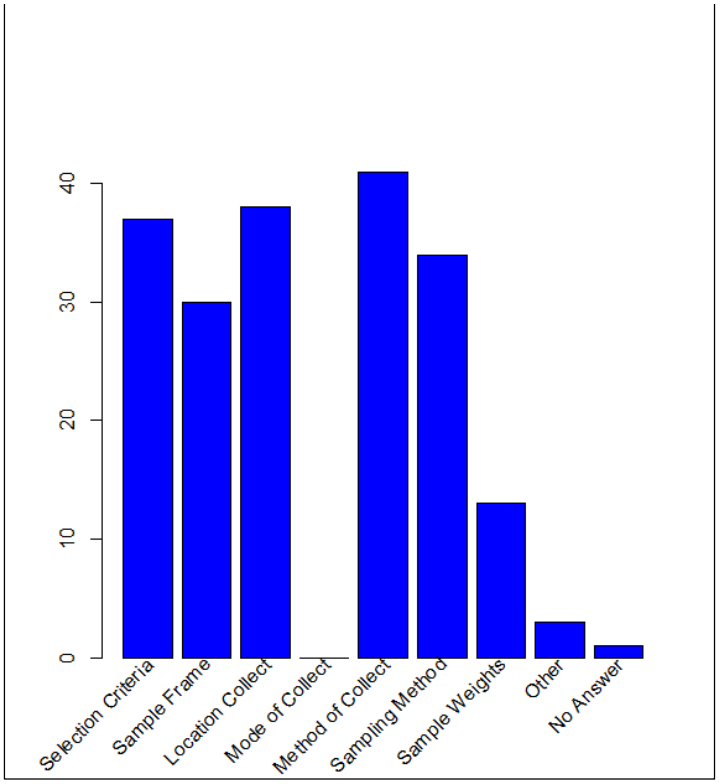

Figure 12: Important Design Elements Bar plot

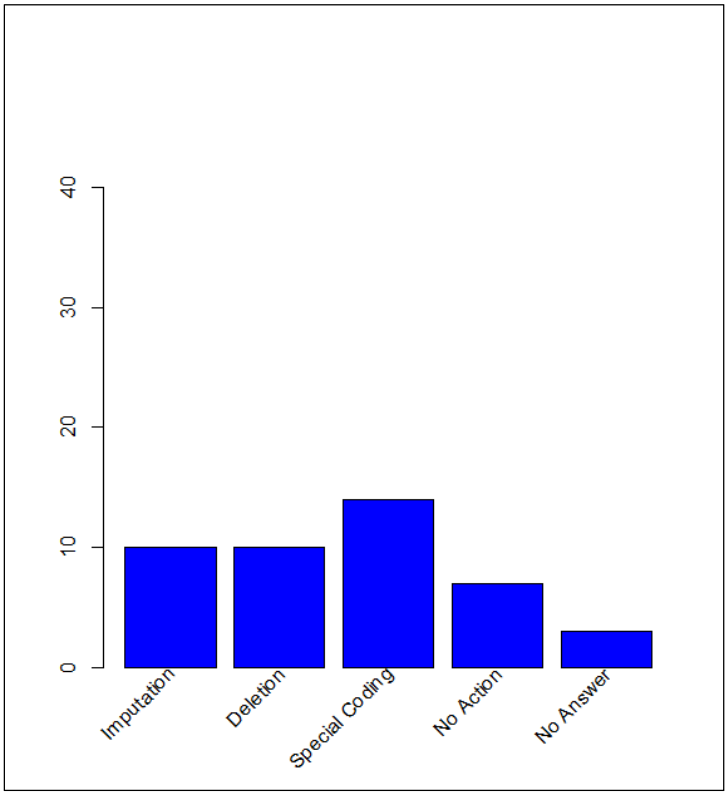

Figure 13: Missing Data Methods Bar plot

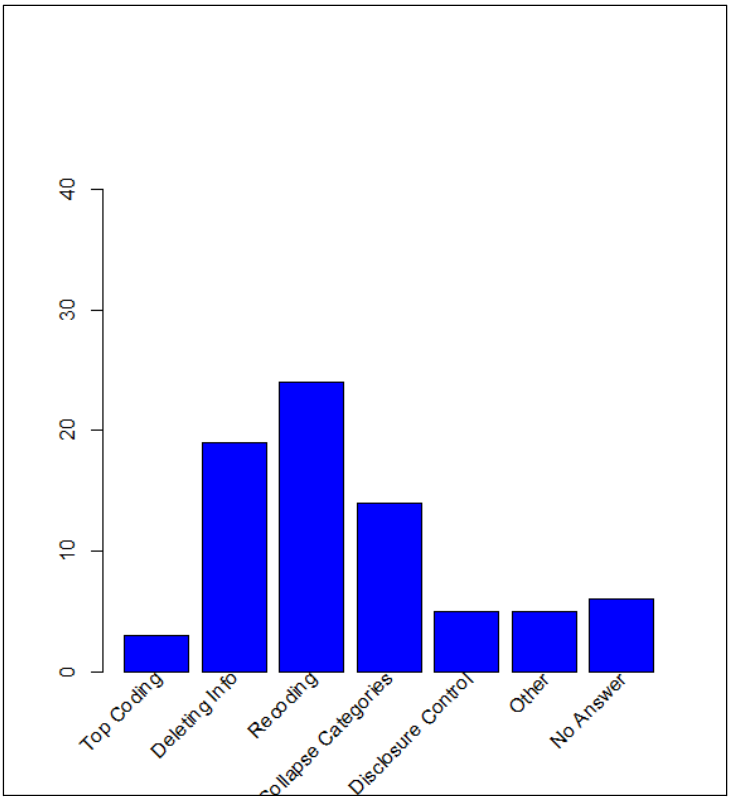

Figure 14: Sensitive Data Methods Bar plot

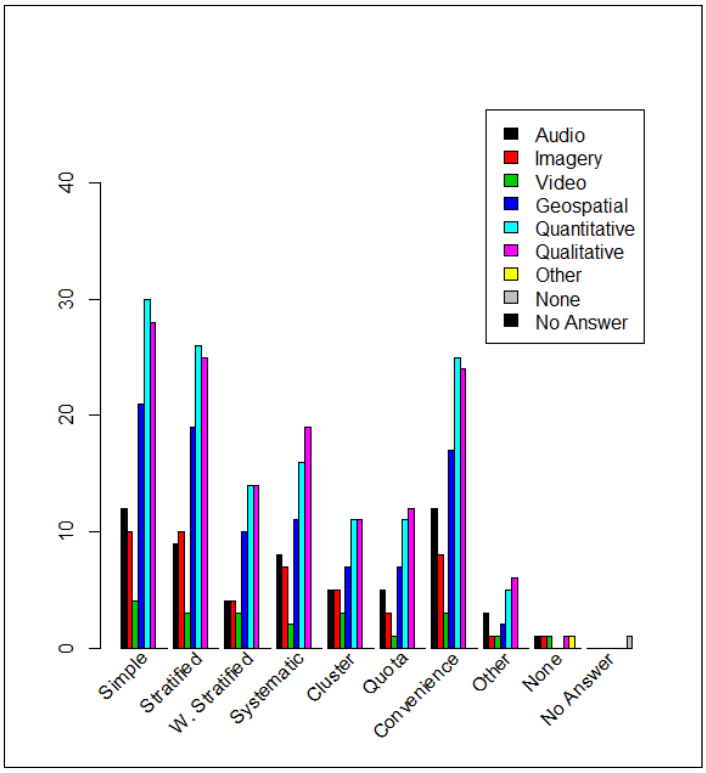

Figure 15: Type of Data vs Sampling Method Bar plot

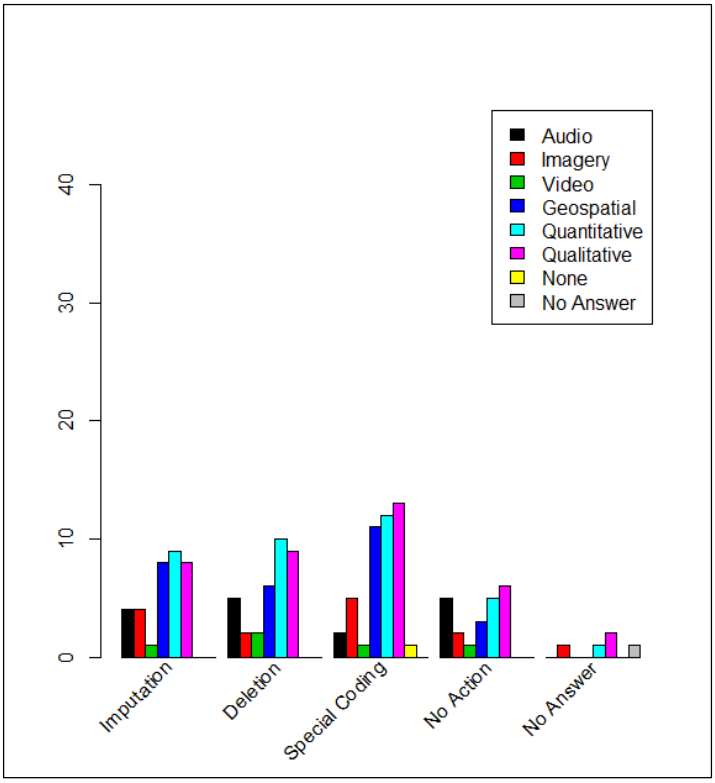

Figure 16: Type of Data vs Missing Data Methods Bar plot

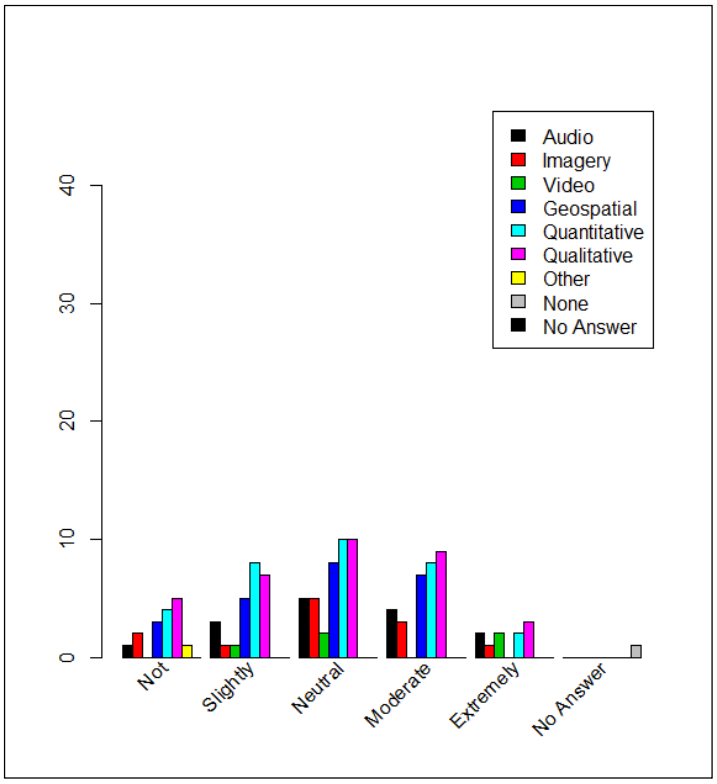

Figure 17: Type of Data vs Sensitivity of Data Bar plot

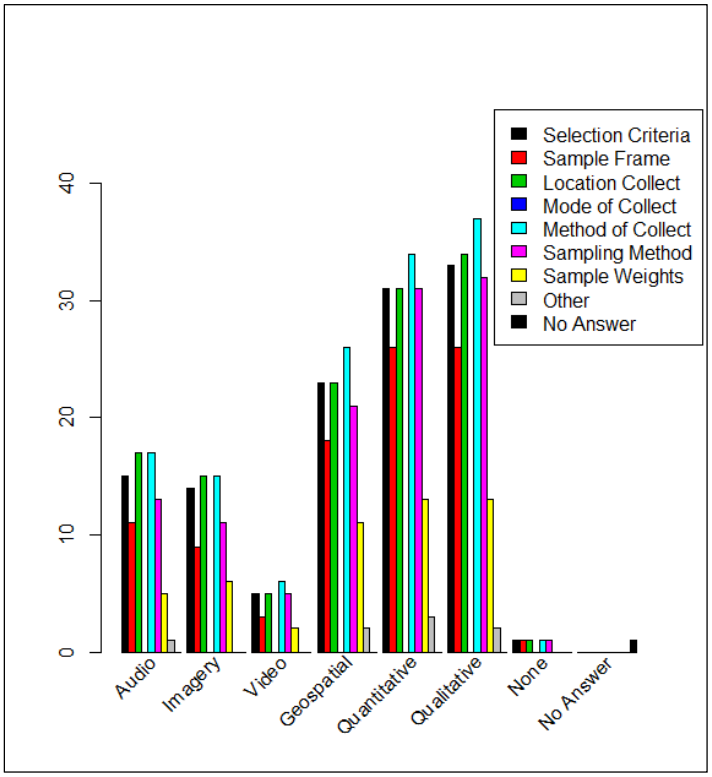

Figure 18: Type of Data vs Design elements Bar plot

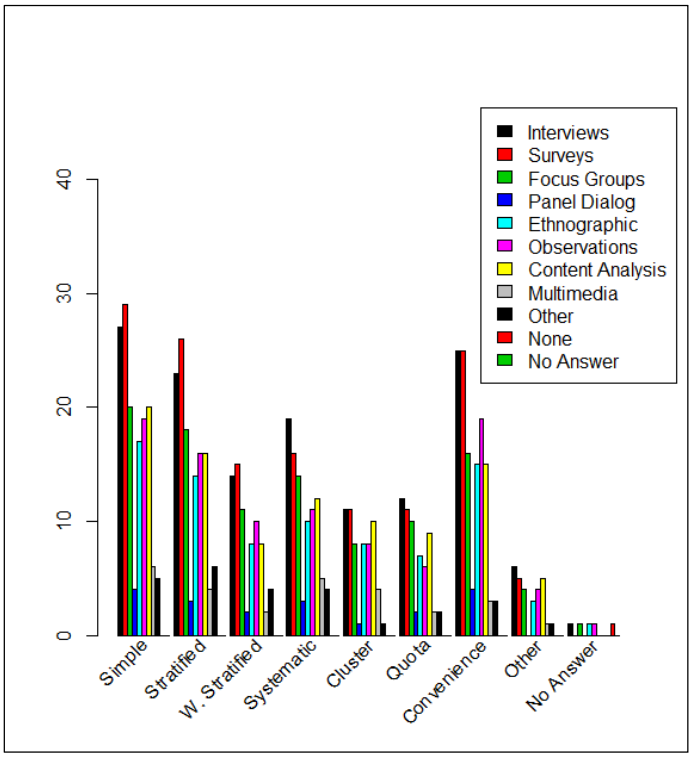

Figure 19: Collection Methods vs Sampling Method Bar plot

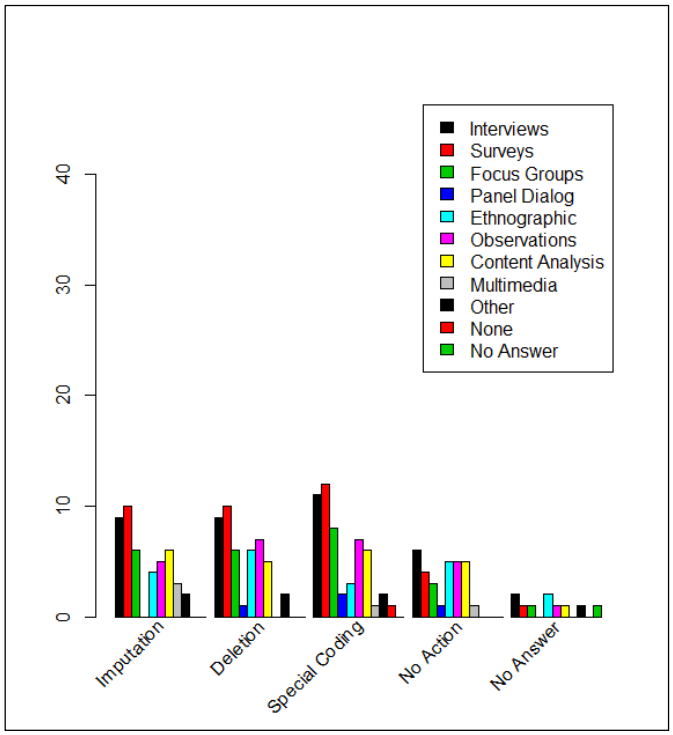

Figure 20: Collection Methods vs Missing Data Methods Bar plot

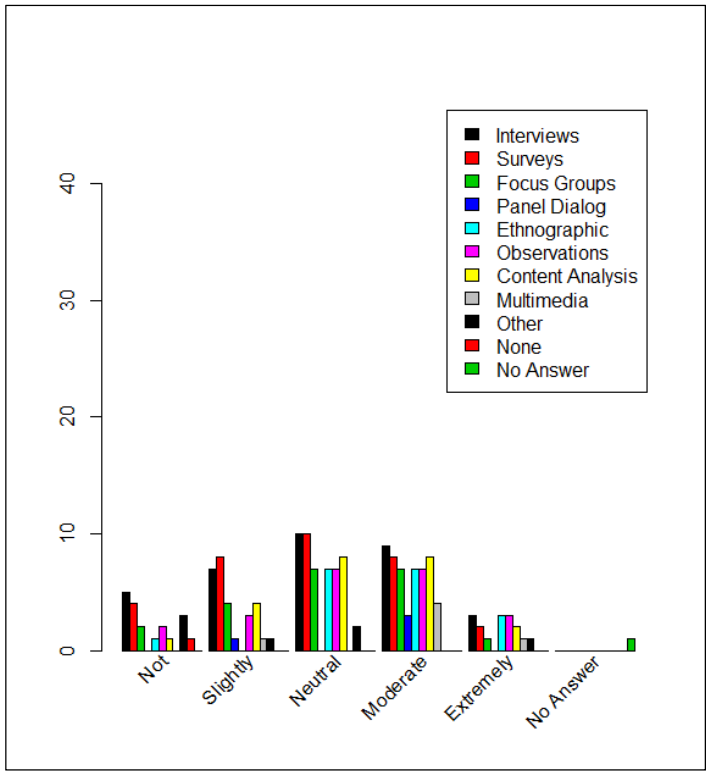

Figure 21: Collection Methods vs Sensitivity of Data Bar plot

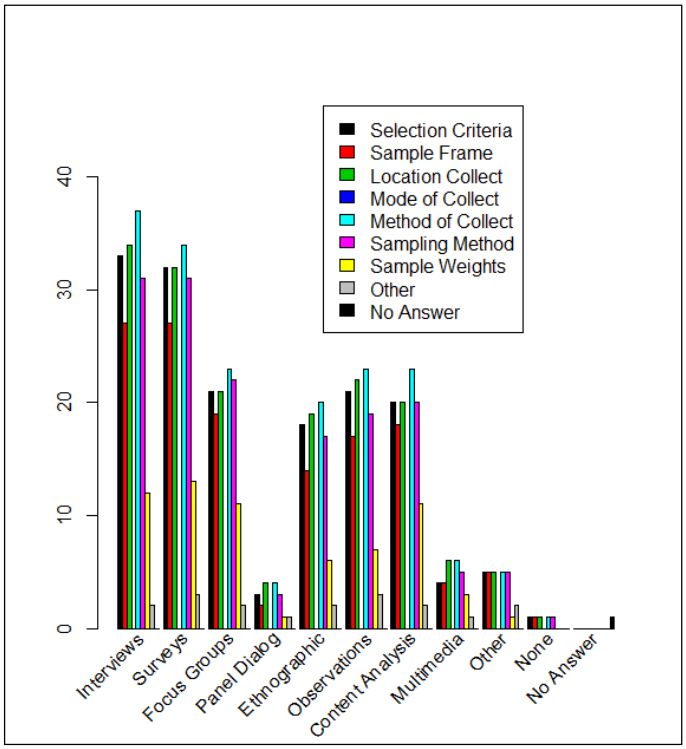

Figure 22: Collection Methods vs Design elements Bar plot

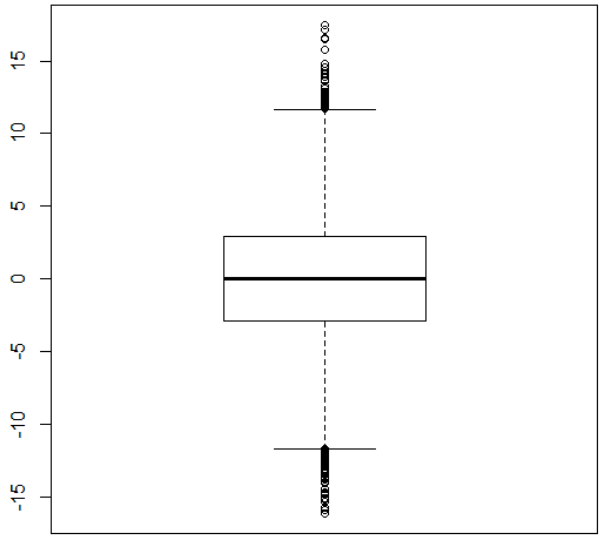

Figure 23: Missing Data Simulation Box plot

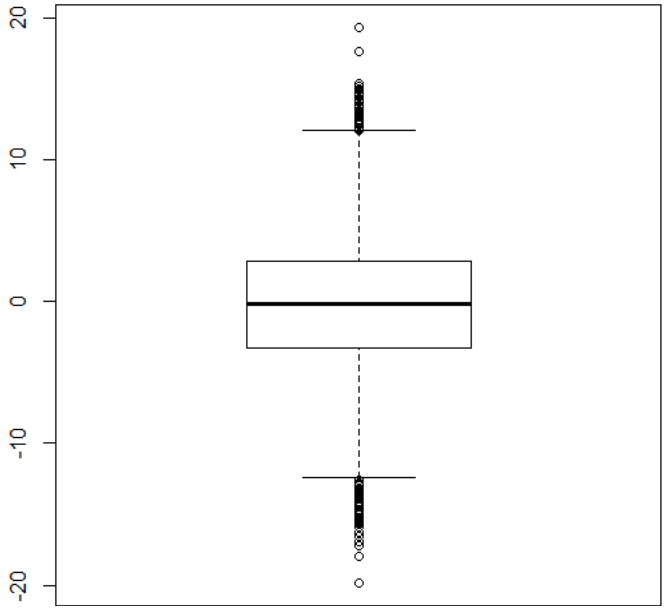

Figure 24: Melting Pot Method 1-1

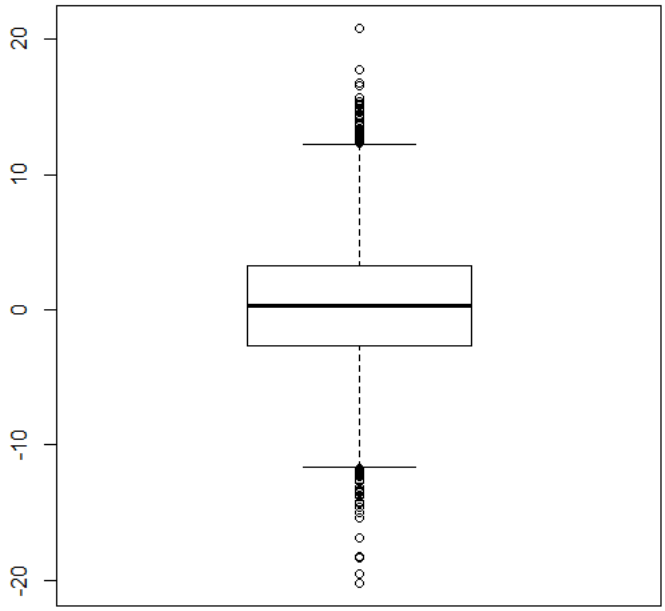

Figure 25: Melting Pot Method 1-2

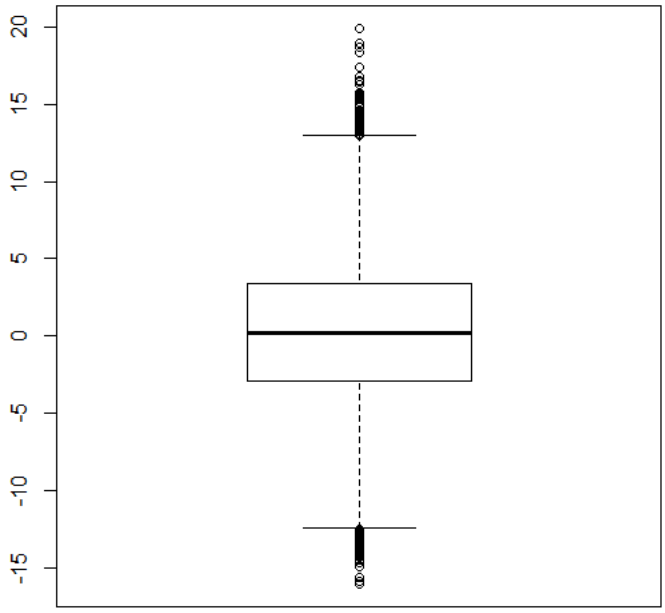

Figure 26: Melting Pot Method 1-3

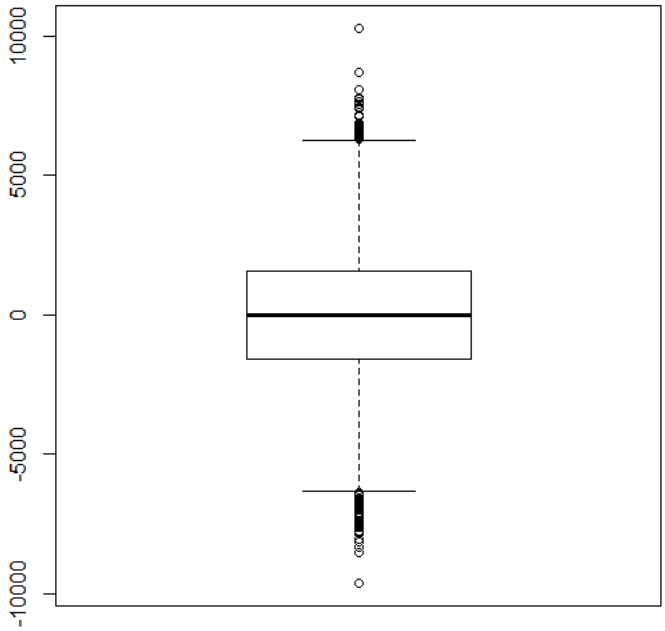

Figure 27: Mean of Means Method 1-1

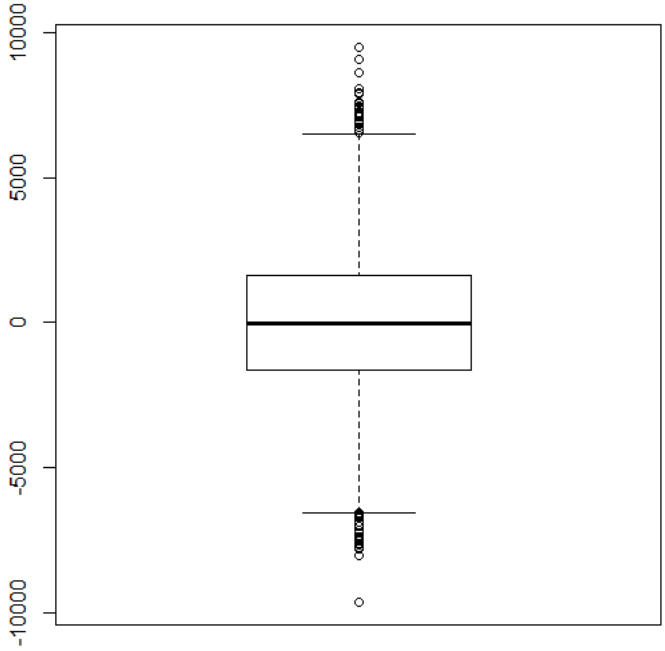

Figure 28: Mean of Means Method 1-2

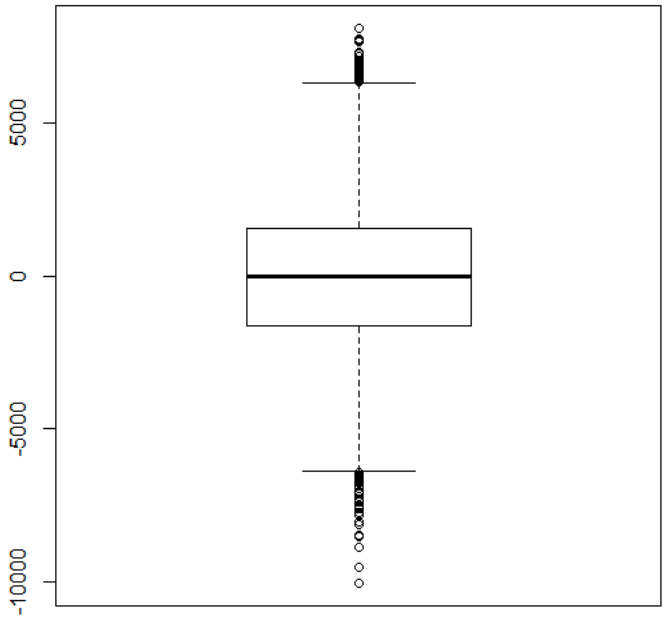

Figure 29: Mean of Means Method 1-3

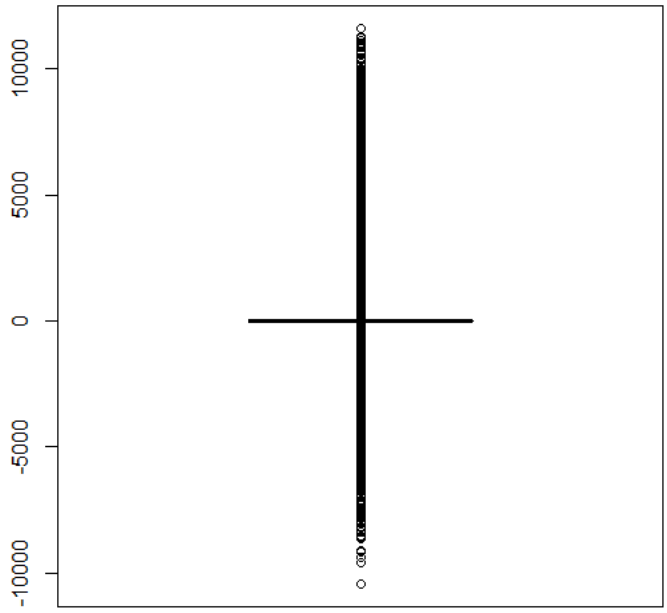

Figure 30: Weight Means Method 1-1

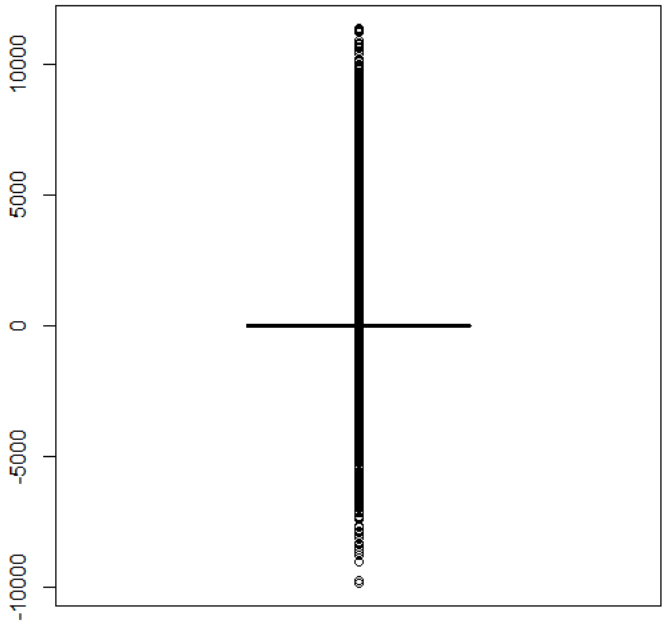

Figure 31: Weighted Means Method 1-2

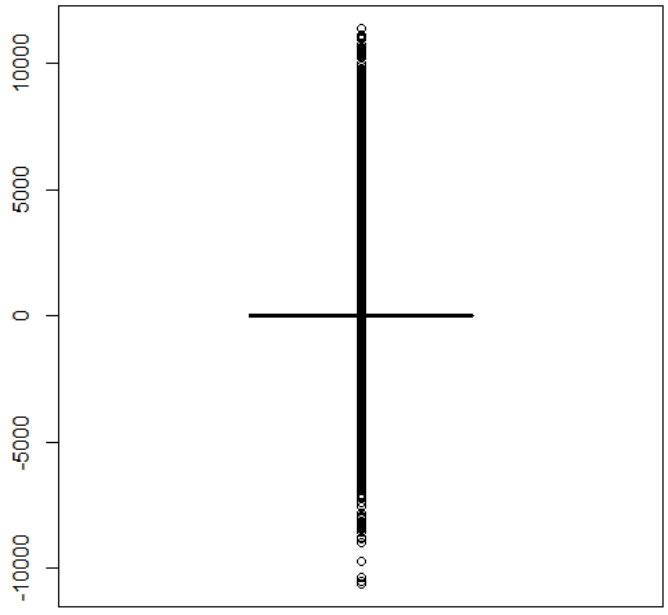

Figure 32: Weighted Means Method 1-3

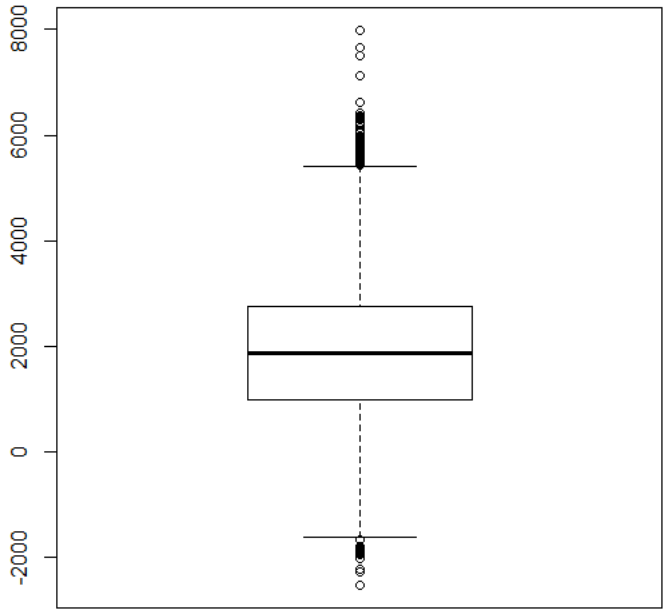

Figure 33: Melting Pot Method 2-1

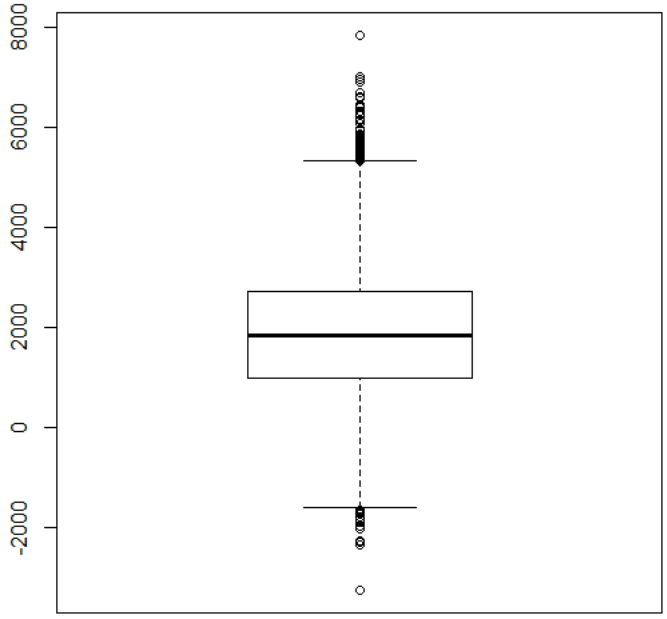

Figure 34: Melting Pot Method 2-2

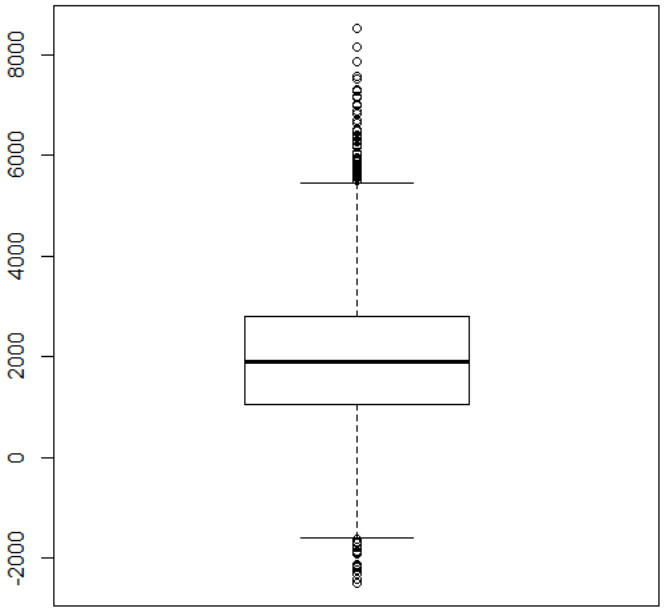

Figure 35: Melting Pot Method 2-3

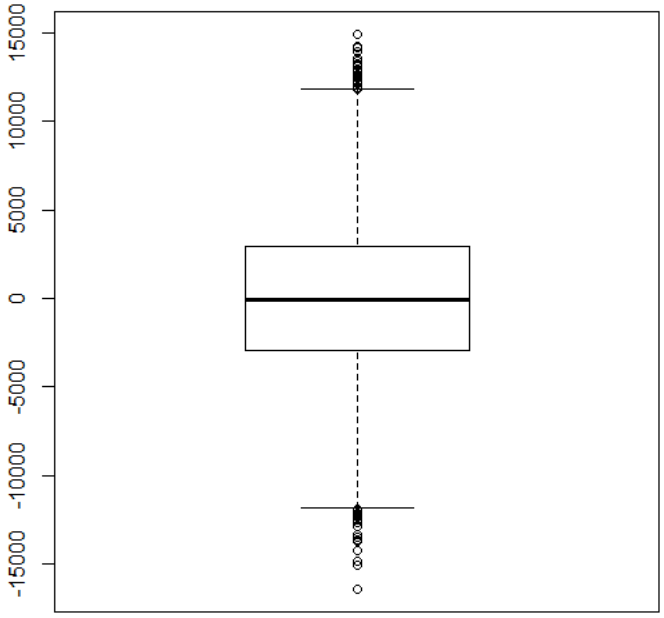

Figure 36: Mean of the Means Method 2-1

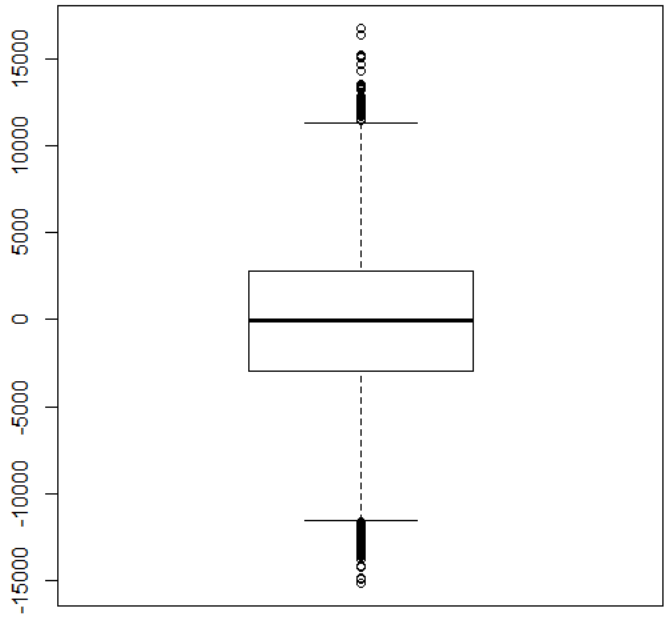

Figure 37: Mean of Means Method 2-2

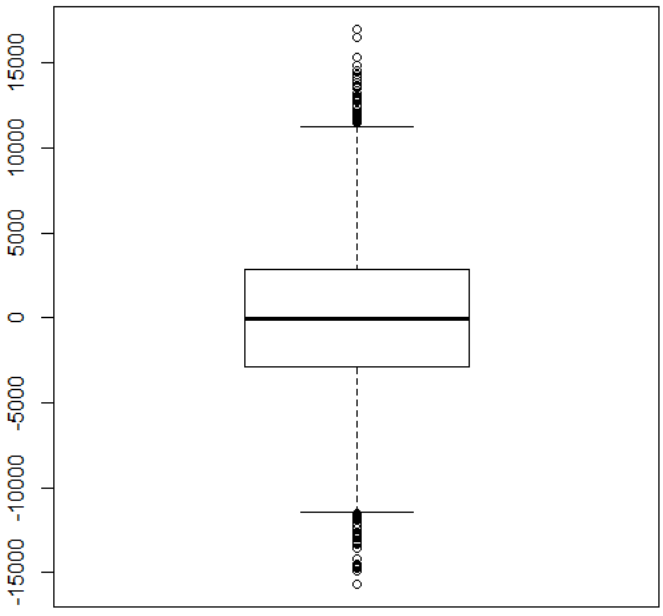

Figure 38: Mean of Means Method 2-3

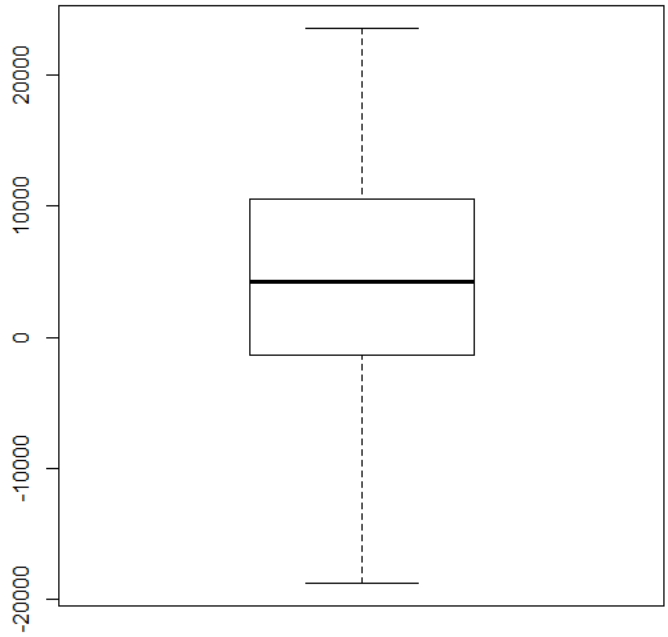

Figure 39: Weighted Means Method 2-1

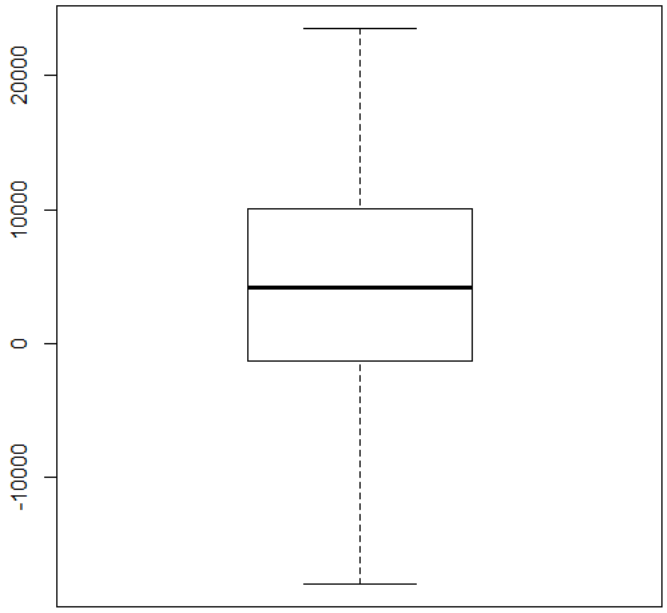

Figure 40: Weighted Means Method 2-2

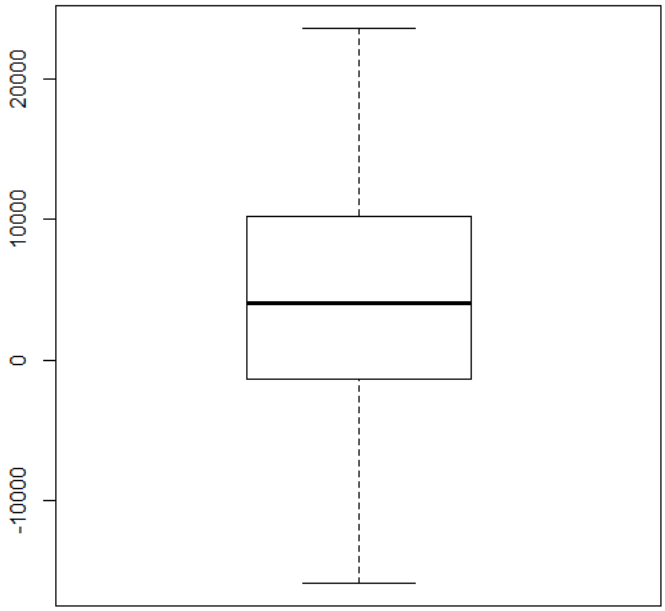

Figure 41: Weighted Means Method 2-3

Appendix B: Tables

Table 1: Importance of Data Sharing in Discipline Frequencies				
Very Important	Moderately important	Neutrally Important	Slightly Important	Not Important
36.6%	22%	17.1%	12.2%	9.8%

Table 2: Importance of Data Sharing Frequencies				
Very Important	Moderately important	Neutrally Important	Slightly Important	Not Important
43.9%	26.8%	7.3%	17.1%	2.4%

Table 3: Importance of Documentation Frequencies				
Very Important	Moderately important	Neutrally Important	Slightly Important	Not Important
73.2%	14.6%	4.9%	4.9%	0.0%

Table 4: Experience with Sharing Data Frequencies				
Always	Often	Sometimes	Rarely	Never
7.3%	29.3%	36.6%	19.5%	4.9%

Table 5: Experience with using Shared Data Frequencies				
Always	Often	Sometimes	Rarely	Never
2.4%	24.5%	39.0%	29.3%	2.4%

Table 6: Type of Data Frequencies						
Audio	Imagery	Video	Geospatial	Quantitative	Qualitative	None
36.2%	31.9%	12.8%	61.7%	80.9%	87.2%	2.1%

Table 7: Collection Method Frequencies				
Interview	Survey	Focus Groups	Panel Dialog	Ethnographic
85.1%	83%	55.3%	10.6%	44.7%
Observations	Content Analysis	Multimedia	Other	None
57.4%	55.5%	12.8%	14.9%	2.1%

Table 8: Sampling Method Frequencies				
Simple Random Sampling	Stratified Random Sampling	Weighted Stratified Sampling	Systematic Random Sampling	Cluster Sampling
66.0%	57.4%	31.9%	40.4%	25.5%
Quota Sampling	Convenience Sampling	Other	None	
25.5%	57.4%	12.8%	4.3%	

Table 9: Important Design Elements Frequencies							
Subject Selection Criteria	Sampling Frame	Location Collected	Mode of Collection	Method of Collection	Sampling Method	Sampling Weights	Other
78.7%	63.8%	80.9%	0.0%	87.2%	72.3%	27.7%	6.4%

Table 10: Frequency of Missing Data Frequencies				
Everytime	Almost Everytime	Occasionally	Almost Never	Never
17.8%	28.9%	42.2%	6.7%	2.2%

Table 11: Sensitivity of Data Frequencies				
Extremely	Moderately	neutral	Slightly	Not at all
7.5%	22.5%	30.0%	20.0%	17.5%

Table 12: Missing Data Methods Frequencies			
Statistical Imputation	Deletion of Records	Special Coding	No Action
21.3%	21.3%	29.8%	14.9%

Table 13: Sensitive Data Methods Frequencies					
Top Coding	Deleting Data	Recoding/ Anonymizing	Collapsing Categories	Disclosure Control	Other
6.4%	40.4%	51.1%	29.8%	10.6%	10.6%

Table 14: Type of Data vs Sampling Method Cross Tabulation					
	Simple Random Sampling	Stratified Random Sampling	Weighted Stratified Sampling	Systematic Random Sampling	Cluster Sampling
Audio	25.5%	19.1%	8.5%	17.0%	10.6%
Imagery	21.3%	21.3%	8.5%	14.9%	10.6%
Video	8.5%	6.4%	6.4%	4.3%	6.4%
Geospatial	44.7%	40.4%	21.3%	23.4%	14.9%
Quantitative	63.8%	55.3%	29.8%	34.0%	23.4%
Qualitative	59.6%	53.2%	29.8%	40.4%	23.4%
None	0.0%	0.0%	0.0%	0.0%	0.0%

Table 14: Type of Data vs Sampling Method Cross Tabulation (Cont.)				
	Quota Sampling	Convenience Sampling	Other	None
Audio	10.6%	25.5%	6.4%	2.1%
Imagery	6.4%	17.0%	2.1%	2.1%
Video	2.1%	6.4%	2.1%	2.1%
Geospatial	14.9%	36.3%	4.3%	0.0%
Quantitative	23.4%	53.2%	10.6%	0.0%
Qualitative	25.5%	51.1%	12.8%	2.1%
None	0.0%	0.0%	0.0%	2.1%

Table 15: Type of Data vs Missing Data Methods Cross Tabulation				
	Statistical Imputation	Deletion of Records	Special Coding	No Action
Audio	8.5%	10.6%	4.3%	10.6%
Imagery	8.5%	4.3%	10.6%	4.3%
Video	2.1%	4.3%	2.1%	2.1%
Geospatial	17.0%	12.8%	23.4%	6.4%
Quantitative	19.1%	21.3%	25.5%	10.6%
Qualitative	17.0%	19.1%	27.7%	12.8%
None	0.0%	0.0%	2.1%	0.0%

Table 16: Type of Data vs Sensitivity of Data Cross Tabulation					
	Extremely	Moderately	Neutral	Slightly	Not at all
Audio	4.3%	8.5%	10.6%	6.4%	2.1%
Imagery	2.1%	6.4%	10.6%	2.1%	4.3%
Video	4.3%	0.0%	4.3%	2.1%	0.0%
Geospatial	0.0%	14.9%	17.0%	10.6%	6.4%
Quantitative	4.3%	17.0%	21.3%	17.0%	8.5%
Qualitative	6.4%	19.1%	21.3%	14.9%	10.6%
None	0.0%	0.0%	0.0%	0.0%	2.1%

Table 17: Type of Data vs Design elements Cross Tabulation				
	Selection Criteria	Sampling Frame	Location of Collection	Mode of Collection
Audio	31.9%	23.4%	36.2%	0.0%
Imagery	29.8%	19.1%	31.9%	0.0%
Video	10.6%	6.4%	10.6%	0.0%
Geospatial	48.9%	38.3%	48.9%	0.0%
Quantitative	66.0%	55.3%	66.0%	0.0%
Qualitative	70.2%	55.3%	72.3%	0.0%
None	2.1%	2.1%	2.1%	0.0%

Table 17: Type of Data vs Design elements Cross Tabulation (Cont.)				
	Method of Collection	Sampling Method	Sampling Weights	Other
Audio	36.2%	27.7%	10.6%	2.1%
Imagery	31.9%	23.4%	12.8%	0.0%
Video	12.8%	10.6%	4.3%	0.0%
Geospatial	55.3%	44.7%	23.4%	4.3%
Quantitative	72.3%	66.0%	27.7%	6.4%
Qualitative	78.7%	68.1%	27.7%	4.3%
None	2.1%	2.1%	0.0%	0.0%

Table 18: Collection Methods vs Sampling Method Cross Tabulation				
	Simple Random Sampling	Stratified Random Sampling	Weighted Stratified Sampling	Systematic Random Sampling
Interview	57.4	48.9%	29.8%	40.4
Survey	61.7%	55.3%	31.9%	34.0%
Focus Group	42.6%	38.3%	23.4%	29.8%
Panel Dialog	8.5%	6.4%	4.3%	6.4%
Ethnographic	36.2%	29.8%	17.0%	21.3%
Observations	40.4%	34.0%	21.3%	23.4%
Content Analysis	42.6%	34.0%	17.0%	25.5%
Multimedia	12.8%	8.5%	4.3%	10.6%
Other	10.6%	12.8%	8.5%	8.5%
None	0.0%	0.0%	0.0%	0.0%

Table 18: Collection Methods vs Sampling Method Cross Tabulation (Cont.)				
	Cluster Sampling	Quota Sampling	Convenience Sampling	Other
Interview	23.4%	25.5%	53.2%	12.8%
Survey	23.4%	23.4%	53.2%	10.6%
Focus Group	17.0%	21.3%	34.0%	8.5%
Panel Dialog	2.1%	4.3%	8.5%	0.0%
Ethnographic	17.0%	14.9%	31.9%	6.4%
Observations	17.0%	12.8%	40.4%	8.5%
Content Analysis	21.3%	19.1%	31.9%	10.6%
Multimedia	8.5%	4.3%	6.4%	2.1%
Other	2.1%	4.3%	6.4%	2.1%
None	0.0%	0.0%	0.0%	0.0%

Table 19: Collection Methods vs Missing Data Methods Cross Tabulation				
	Statistical Imputation	Deletion of Records	Special Coding	No Action
Interview	19.1%	19.1%	23.4	8.5%
Survey	21.3%	21.3%	25.5%	12.8%
Focus Group	12.8%	12.8%	17.0%	6.4%
Panel Dialog	0.0%	2.1%	4.3%	2.1%
Ethnographic	8.5%	12.8%	6.4%	10.6%
Observations	10.6%	14.9%	14.9%	10.6%
Content Analysis	12.8%	10.6%	12.8%	10.6%
Multimedia	6.4%	0.0%	2.1%	2.1%
Other	4.3%	4.3%	4.3%	0.0%
None	0.0%	0.0%	2.1%	0.0%

Table 20: Collection Methods vs Sensitivity of Data Cross Tabulation					
	Extremely	Moderate	Neutral	Slightly	Not at all
Interview	6.3%	19.1%	21.3%	14.9%	10.6%
Survey	4.3%	17.0%	21.3%	17.0%	8.5%
Focus Group	2.1%	14.9%	14.9%	8.5%	4.3%
Panel Dialog	0.0%	6.4%	0.0%	2.1%	0.0%
Ethnographic	6.4%	14.9%	14.9%	0.0%	2.1%
Observations	6.4%	14.9%	14.9%	6.4%	4.3%
Content Analysis	4.3%	17.0%	17.0%	8.5%	2.1%
Multimedia	2.1%	8.5%	0.0%	2.1%	0.0%
Other	2.1%	0.0%	4.3%	2.1%	6.4%
None	0.0%	0.0%	0.0%	0.0%	2.1%

Table 21: Collection Methods vs Design elements Cross Tabulation				
	Selection Criteria	Sample Frame	Location of Collection	Mode of Collection
Interview	70.2%	57.4%	72.3%	0.0%
Survey	68.1%	57.4%	68.1%	0.0%
Focus Group	44.7%	40.4%	44.7%	0.0%
Panel Dialog	6.4%	4.3%	8.5%	0.0%
Ethnographic	38.3%	29.8%	40.4%	0.0%
Observations	44.7%	36.2%	46.8%	0.0%
Content Analysis	42.6%	38.3%	42.6%	0.0%
Multimedia	8.5%	8.5%	12.8%	0.0%
Other	10.6%	10.6%	10.6%	0.0%
None	2.1%	2.1%	2.1%	0.0%

Table 21: Collection Methods vs Design elements Cross Tabulation (Cont.)				
	Method of Collection	Sampling Method	Sample Weights	Other
Interview	78.7%	66.0%	25.5%	4.3%
Survey	72.3%	66.0%	27.7%	6.4%
Focus Group	48.9%	46.8%	23.4%	4.3%
Panel Dialog	8.5%	6.4%	2.1%	2.1%
Ethnographic	42.6%	36.2%	12.8%	4.3%
Observations	48.9%	40.4%	14.9%	6.4%
Content Analysis	48.9%	42.6%	23.4%	4.3%
Multimedia	12.8%	10.6%	6.4%	2.1%
Other	10.6%	10.6%	2.1%	4.3%
None	2.1%	2.1%	0.0%	0.0%

Table 22: Simulation Results Population 1									
	Melting Pot Method			Mean of the Means Method			Weighted Means Method		
	Trail 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3
Coverage	92.9%	96.0%	97.3%	100%	100%	100%	95.7%	95.3%	96.0%
Bias	-0.156	0.256	0.206	-39.7	10.4	-0.429	3.65	3.34	3.40
Average Error	9.22	9.23	9.22	2.34e5	2.31e5	2.31e5	1.87e3	1.76e3	1.75e3

Table 23: Simulation Results Population 2									
	Method 1			Method 2			Method 3		
	Trail 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3
Coverage	16.4%	20.4%	20.3%	100%	100%	100%	72.9%	73.2%	72.0%
Bias	1.86e3	1.84e3	1.8e3	-43.5	-18.5	-55.8	4.06e3	4.20e3	4.24e3
Average Error	821.9	821.5	820.5	4.19e5	4.21e5	4.20e5	1.11e4	1.11e4	1.11e4

Bibliography

Bechhofer, F., & Paterson, L. (2000). *Principles of Research Design in the Social Sciences*. London: Routledge.

Bloom, H. (2008). The Core Analytics of Randomized Experiments for Social Research. In P. Alasuutari, L. Bickman, and J. Brannen (Eds.), *The SAGE Handbook of Social Research Methods*. http://dx.doi.org/10.4135/9781446212165

Corti, L. (2008). Data Management. In L. M. Given (Ed.), *The Sage Encyclopedia of Qualitative Research Methods*. http://dx.doi.org/10.4135/9781412963909.n98

Corti, L. (2008). Data Storage. . In L. M. Given (Ed.), *The Sage Encyclopedia of Qualitative Research Methods*. http://dx.doi.org/10.4135/9781412963909.n101

Corti, L. (2008). Secondary Analysis. . In L. M. Given (Ed.), *The Sage Encyclopedia of Qualitative Research Methods*. http://dx.doi.org/10.4135/9781412963909.n415

Cragin, M. H., Palmer, C. L., Carlson, J. R., and Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023-4038. Retrieved from http://www.jstor.org/stable/25704697

Daly, M. (2003). Methodology. In R. L. Miller, and J. D. Brewer (Eds.), *The A-Z of Social Research*. http://ida.lib.uidaho.edu:4101/10.4135/9780857020024

Greenstein, T. N. (2006). Using Other People's Data. In *Methods of Family Research,* 2nd ed. http://dx.doi.org/10.4135/9781412990233.d14

Hakim, C. (1982). Secondary Analysis and the Relationship between Official and Academic Social Research. *Sociology*, *16*(1). 12-28. Retrieved from http://www.jstor.org/stable/42852390

Higgins, J. P., and Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*, Version 5.1.0. Retrieved from www.cochrane-handbook.org.

Holdren, J. P. (2013, February 22). *Increasing Access to the Results of Federally Funded Scientific Research*. [Memorandum]. Retrieved from: https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Ioannidis, J. P., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., ... and Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *The Lancet*, *383*(9912), 166-175. http://dx.doi.org/10.1016/S0140-6736(13)62227-8

Litwin, M. S. (1995). Creating and using a codebook. In *The Survey Kit: How to measure survey reliability and validity*. http://ida.lib.uidaho.edu:4101/10.4135/9781483348957.n5

Macdonald, A. (2006). Digital archiving, curation and corporate objectives in pharmaceuticals. *Journal of Medical Marketing*, *6*(2), 115-118. http://dx.doi.org/10.1057/palgrave.jmm.5050030

Miller, R. (2003). Research Design. In R. L., Miller, and J. D. Brewer (Eds.), *The A-Z of Social Research.* http://ida.lib.uidaho.edu:4101/10.4135/9780857020024

Schwartz, C. E., Ahmed, S., Sawatzky, R., Sajobi, T., Mayo, N., Finkelstein, J., ... and Sprangers, M. A. (2013). Guidelines for secondary analysis in search of response shift. *Quality of Life Research*, *22*(10). 2663-2673. http://ida.lib.uidaho.edu:2432/docview/1465144640?OpenUrlRefId=info:xri/sid:primo&accountid=14551

Tanenbaum, E. (1980). Secondary Analysis, Data Banks and Geography. *Area*, *12*(1). 33-35. Retrieved from http://www.jstor.org/stable/20001530

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. http://ida.lib.uidaho.edu:4101/10.4135/9781412984980

Yen, J. (2002). *Combining studies*. Retrieved from http://www.nist.gov/itl/sed/training/upload/combine-1.pdf