Working in Data Mines: Conducting Multiple Analyses on Qualitative Data Sets

Dr. Deirdre-Annaliese Nicole Hunter, Virginia Tech

Dr. Deirdre Hunter conducts engineering education research at Virginia Tech and is the Director of U.S. Development at La Gran Familia de Gregory in Chihuahua, Mexico. Her current research is in the areas of problem-based learning facilitation and teaching metacognition. Her research strengths include research design and implementation using qualitative methods. She has a Ph.D. in Engineering Education from Virginia Tech, a B.S. in Mechanical Engineering from Syracuse University, and a A.S. in Engineering Science from Onondaga Community College, NY.

Dr. Philip Reid Brown, Virginia Tech

Philip Brown recently received his PhD from the Department of Engineering Education at Virgnia Tech. His research interests include the use of motivation, cognition and learning theories in engineering education research and practice, and better understanding student perspectives in engineering programs.

Working in Data Mines: Conducting Multiple Analyses on Qualitative Data Sets

Abstract: This paper has two purposes: to introduce the idea of mining qualitative data to new engineering education researchers, and to provide useful examples for reference. While it is common to see large quantitative data sets being mined for new findings, large qualitative data sets (whether interviews or observations) are often only used for one research agenda. Qualitative data sets—like their quantitative counterparts—are rich enough in information to support secondary analyses, and researchers should consider them as viable sources to support multiple investigative agendas.

There are multiple reasons for re-using qualitative data sets. Large qualitative data sets require significant time and resources for data collection and transcription. Particularly for faculty that face limited funding, and graduate students that face limited timelines for their theses and dissertations, pre-existing large qualitative data sets are valuable resources that can reduce the time and resources necessary for traditional qualitative studies. However, secondary analysis of qualitative data (i.e., analysis of an existing qualitative data set to answer new research questions) comes with challenges. For example, the purpose of the new analysis can be quite different from that of the original study, which can raise questions of alignment of the data with the research questions and the data analysis methods.

In this paper, we present an example of how researchers used one set of qualitative data to support multiple research agendas. First, we describe the large longitudinal qualitative data set, including the data collection methods and the original research questions guiding the study. Subsequently, we discuss how three graduate students separately used this qualitative data set in completion of their dissertations. Specifically, two graduate students used the data for secondary analyses to ask new questions of the data and another graduate student used the data to pilot an observation protocol as part of her dissertation study. We discuss how those graduate students determined that the data set was viable for their individual research agendas, and outline some guidelines that others could use in doing so. Lastly, we conclude with the implications particular in designing a qualitative study around secondary data analysis. We believe that the information in this paper is valuable to graduate students and new faculty considering new research avenues with limited resources, in an effort to maximize the usefulness of previously dedicated resources.

Introduction

In the worlds of statistics and quantitative research, the term 'data mining' is used to refer to techniques that help researchers investigate new findings from existing data sets[1, 2]**.** Data Mining usually involves analyzing extremely large databases, but the methods involved are applied to smaller sets of data as well. In all applications, the purpose of data mining is to develop new knowledge from existing information.

In qualitative research, we often use the term secondary analysis for the repurposing of existing data for new investigations[3-5]. Broadly, this includes any subsequent use of a qualitative data set for research purposes beyond those for which it was originally collected, be it by the original

researcher or other entities. More specifically, Corti and Thompson[6] have listed five uses for secondary analysis:

1) Historical description through primary sources
2) Follow-up work for the original study
3) Re-analysis for a new purpose
4) Research design and methodological advancement.
5) Verification of original results

In this paper, we discuss a more specific case of secondary analysis that we refer to as qualitative data mining. Qualitative data mining is the process of finding datasets for purposes of secondary analysis, and matching data to your research purposes. In this paper, we overview qualitative data mining as a viable way for graduate students in engineering education to explore their own research avenues. Our examples of qualitative data mining fit most closely with the third and fourth uses of secondary analysis listed by Corti and Thompson.

Note that, in this paper, we are not discussing quantitative and statistical data mining techniques applied to qualitative data, like textual data mining[7]. All of the techniques we describe are qualitative in nature.

Searching for existing qualitative data sets is not a new idea. Seale[3] notes that archives of qualitative data have existed for decades. However, while engineering education research utilizing secondary analysis exists[8-12], an overview for how graduate students and other individuals might mine existing qualitative data does not exist. In this paper, we present such an overview. We provide the example of how one existing qualitative data set was used for three dissertation studies. We show how each study had its own considerations for the use of that data set, and how those considerations could help others identify whether a data set is useful for them while they are mining qualitative data.

There are two main reasons that researchers should consider qualitative data mining. Seale[3] notes that secondary analyses like those that are performed through qualitative data mining are beneficial to both the researcher and the research community in general, stating that, "It is both economical and ethical to analyze existing data sets rather than collecting new ones, if these data sets are adequate to answer the research questions being asked." (p. 351). Economical refers to time and resources that researchers save. Ethical refers to respecting the time and resources of others required in collecting large data sets. The former point is particularly relevant to graduate students, as they are typically more constrained by both time and resources in their research agendas.

Background

Here we present how three researchers used one large data set (that included both qualitative and quantitative data) to meet their own qualitative research needs. First, we describe the original research project associated with this data set, as well as the breadth of data within the data set. The original project was a three year NSF-funded project focused on how different components of problem-based learning (PBL) pedagogy impact women's interest and persistence in

engineering[13, 14]. The PBL project was a mixed-methods, two-site, longitudinal investigation that included classroom observations and video recordings of student teams, audio recordings of instructor training, semi-structured student interviews, and quantitative student surveys. Data collection spanned all three years. The breadth, depth, and longitudinal span of the data for this project were important factors when considering secondary analysis of this data. Year 1 of the project included students enrolled in courses at both sites. These students (Cohort 1 and 2) responded to surveys, and participated in observations of their team interactions and in semi-structured interviews. In Year 2, the Cohorts 1 and 2 from Year 1 participated in follow-up interviews and responded to follow-up surveys. Also in Year 2, Cohort 3 participated in a new classroom implementation and were surveyed, observed, and interviewed. In addition to the student data for Cohort 3, the facilitator training sessions were also recorded. In year 3 of the project, students from all three cohorts were again surveyed and interviewed. All data collected during the three-year project is summarized in Table 1, and is referred to as the PBL data set.

Table 1 - Data Collected for the NSF-funded project (PBL data set).

Context		Data Type	Data Collection Year		
			Year 1	Year 2	Year 3
Site 1	Cohort 1	Classroom Observations	√	-	-
		Student Interviews & Surveys	√	√	√
Site 2	Cohort 2	Classroom Observations	√	-	-
		Student Interviews & Surveys	√	√	√
	Cohort 3	Classroom Observations	-	√	-
		Facilitator Training Observation	-	√	-
		Student Interviews & Surveys	-	√	√

Key: √ data collected; - data not collected

In addition to the range of data types collected and the longitudinal perspective, there were a large number of participants across the years and extensive contact hours during the observations and interviews. More than 1,300 individual student participated throughout all phases of the project. The observations included nearly 200 students with approximately 350 hours of observed interactions. For Cohort 1 there were five teams of students, where each team was observed for approximately nine hours spread over six weeks. In Cohort 2, six teams of students were observed for over nine hours each over five weeks. Lastly, in Cohort 3, 24 teams were observed for over nine hours each over five weeks. Each interview lasted between 45 minutes and 1 hour and included 56 different student participants, with a total of 104 interviews. Additionally, a total of 3,177 were collected over the three year project. These numbers are summarized in Table 2.

Table 2 – Number of Student Participants in Each Phase of Data Collection

	Classroom Observations		Interviews			Surveys		
	Year 1	Year 2	Year 1	Year 2	Year 3	Year 1	Year2	Year3
Cohort 1	48		19	11	10	64 pre 35 post	26	26
Cohort 2	30		14	8	8	476 pre 374 post	177	170
Cohort 3		120	-	23	11	-	843 pre 715 post	271

Qualitative Data Mining Examples

We present three studies that serve as examples of how to mine secondary data for new research purposes. In particular the first is an example of using secondary data for *research design or methodological advancement*, where and the second and third are examples of *re-analysis* of secondary data for a new research purpose.

Example 1 – Data Mining for Research Design and Methodological Advancement

In our first example, one graduate student used the classroom observation data to aid the development of an observation protocol as a research tool for her dissertation study. The focus of her dissertation was on studying engineering students' use of metacognition in naturalistic settings[15]. While there had been prior studies of metacognition use in other populations, almost no prior work existed for studying metacognition in naturalistic settings, particularly with engineering students. To achieve this research aim, this graduate student needed to develop a new observation method or protocol. To do this, she first developed a draft protocol based on prior literature and conversations with other researchers in the field. Next, to test the draft protocol she used the video-recorded classroom observations. She watched the video-recorded classroom observations and followed the draft observation protocol as if the observations were happening live. In doing this, she was able to determine if observing students interactions and the guidance of the protocol would enable her to observe the metacognition phenomenon she intended to study. After this test of the draft protocol, she was able to make adjustments in the protocol; the test also served as practice for her in learning her role as an observer. In this example, the graduate student was able to advance her research methods by further developing the observation protocol, prior to collecting new data. Though there was no re-analysis of data from the PBL data set, the student still needed to consider whether the data set would meet the needs of her work. Specifically, she needed to consider the context of the observations and if there was a close enough match to the context of her study. Additionally, she had to consider the form in which the data was stored. In the original PBL data set, the classroom observations were video recorded, along with the collection of field notes, and though the observations were transcribed, the videos remained intact and stored with the data set. Having the original videos was critical to being able to develop the research tool necessary for this dissertation study.

Example 2 – Re-analysis of Observations for a New Research Purpose

In our second example, another graduate student re-analyzed the classroom observation data, along with the facilitator training observations to answer new research questions for her dissertation. The intention of the original data collection was to analyze student motivation; however, this graduate student re-analyzed the observations with the purpose of understanding instructor roles, particularly PBL facilitation strategies enacted by instructors in introductory engineering courses[10].

Though there was a shift in the focus between the original study and the secondary study, the context between the new research intentions for the secondary study remained consistent with the original study context. Both studies were interested in behaviors or attitudes of participants in a first-year engineering PBL course. Additionally, since the observations were recorded (both audio and video) and transcribed verbatim the collected data was not biased by a particular theoretical framework and did not cause a problem concerning the difference in focus. Additionally, because the focus on the secondary study was on the facilitators, Cohort 1 was chosen for analysis since the facilitators remained with the group throughout the time, which was not the case for the other cohorts. This was an important piece of information available to the graduate student because of her participation in larger grant. A benefit to having videos of observations is that they can be analyzed for any number of research inquiries as long as care was taken to ensure consistency with the context of the larger project. While in this case the context under which the data were collected is inescapable, the data collection methods and instruments had little impact on biasing the video data.

This graduate student researcher had the benefit of helping to collect and analyze some of the data for the original research study, and thus was familiar with the data set, the participants, and the context. This familiarity helped her to understand both the benefits and the limitations of using this data set for her dissertation study. Among the benefits, she knew how rich the data set was and that the video recordings and transcripts of the observations would be ample enough to provide insight into the instructor behaviors. Additionally, being close to the collection of the data she was more aware of potential biases in the data and her own biases. For example, she had the opportunity to help conduct the student follow-up interviews and talk with administrators at the research sites; this opportunity helped her to understand the culture and expectations of the learning environments at each research site. As far as limitations, the original study was focused on student behaviors and attitudes, therefore, the interviews were conducted only with students. In the secondary analysis with the focus on instructors' behaviors, interviews with the instructors may have been helpful to understand the intention of instructors' behaviors.

Example 3 – Re-analysis of Interviews for a New Research Purpose

In our third example, we present how another graduate student identified the interview data from the PBL data set as useful for his research needs and utilized it as a central component of his dissertation research. The researcher was investigating the career goals that undergraduate students have, and how those students use their career goals to make choices in the present[16]. In choosing a data set for this research question, the researcher had to consider whether a

perspective data set contained students discussing their post-graduation career goals, and whether that discussion included connections to academic choices that students made.

The researcher decided on interview data as the ideal kind of data set for analyzing this question. Interview data is advantageous in that it asks students directed questions, and can lead to insights that could not be obtained from observations. For secondary analysis, however, interview data presents some additional challenges. In addition to a need for a match in context, as in the previous example, one must consider whether the interview protocol asks questions relevant to desired research questions, and whether the ensuing discussion produced by the implementation of that protocol is sufficient for the research purposes. The PBL data set included a selection of interviews with students conducted after they had completed one of the first-year courses that were the focus of the original study. Students were interviewed at the end of their first year, while a subset of those students were interviewed in their second and third years, as well.

In this case, the researcher had previous experience with the data set as a research assistant, analyzing it for some of its original research purposes, and another secondary analysis that resulted in a conference paper[12]. The context of the data set matched the context of his research question: both involved students enrolled in undergraduate engineering programs. He had some recollection that, in some of the interviews, there was some discussion of post-graduation career goals and choices that students made. He then examined the interview protocol for evidence that each interview would contain prompts for similar discussion. In doing so, the researcher identified relevant topics for the exploration of his research question. Next, he looked into individual interviews to determine whether these prompts produced sufficient discussion to warrant analysis. Some things he looked for were:

- Whether interviewers asked follow-up questions to brief or incomplete student responses.
- Whether the prompts to each student were worded consistently.
- Whether the prompts were interpreted consistently by each student.
- The amount of detail students went into.

After considering the above-stated factors, and the number of interviews available to him (57, in total), the researcher decided that the data set was viable for use in his study. In deciding this, he acknowledged some limitations of its use. The inability to ask pointed follow-up questions meant that there were avenues of inquiry that would have been useful for deeper exploration into the research question that the researcher was nonetheless unable to explore. However, these limitations were mitigated by the advantages of this data set. It included an abundant number of interviews, included a sufficient amount of relevant discussion, and had transcripts that were immediately available for analysis. This allowed the researcher to explore a research avenue that would have otherwise taken a drastically increased amount of time, and resources to complete.

Considerations for Qualitative Data Mining

From the examples above, we suggest five considerations for researchers who are interested in qualitative data mining. These guidelines are adapted from other literature on secondary analysis, and have been expanded for the particular considerations of qualitative data mining. These considerations are:

1. What is your research question?
2. What is the context of a dataset?
3. How well will a dataset help you answer your research question?
4. What kind of access do you have to a dataset?
5. Are there any other logistical considerations associated with a dataset?

What is your research question?

The first consideration in qualitative data mining is not exclusive to secondary analysis, or even qualitative research. Formulating a research question is central to investigative research of all kinds, though less central to methodological papers like this[17]. While the process of qualitative data mining does not require you to start with a research question, a researcher should consider whether a possible research question is suitable for secondary analysis. If you have a research question in mind, consider whether it requires you to have access to a specific population or location. You can begin searching for data sets with those factors in mind.

In the examples above, the researchers had research questions in mind prior to considering the PBL data set for use in their study. This helped them evaluate whether the dataset was appropriate for their individual studies. As a specific example, the third researcher was able to match prompts in interview protocols to his research question to see if appropriate types of inquiries were made for the purposes of his study.

If you are considering datasets prior to fully formulating a research question, consider three things: 1) What is it about particular data sets that interests you?, 2) Is there information contained in the dataset that was not explored in the original study?, and 3) What kinds of findings you could draw from the data that would be of interest to the stakeholders of a study you would want to conduct? These considerations will help you formulate a research question.

What is the context of a dataset?

When looking at a dataset for secondary analysis, the first thing to consider is the context of the original data set. The secondary researcher needs to first develop an understanding of the context of the original data, including why the data was collected, and what information the data might contain. In the above examples, we noted how context was an important consideration for each researcher in matching the dataset to their research needs. Context can be established both by knowing the population studied within a data collection process, the original research questions explored in data collection, the story behind the collection of the data set, any protocols for data collection, and any other contextual or supplementary information[18-20]. All of this information will help you in considering the viability of a data set for your research purposes.

As an example of the usefulness of developing understanding of context of the original data set, we refer to the second study discussed in the examples sections. The researcher knew that observation data was available from two sites. However, due to the contextual understanding she had gained from being familiar with that data set, she knew that the observations at University 1 were better suited for studying faculty behavior due to camera placement differences at each site and the ability to see the interactions between the facilitator and student groups.

There are a few ways that a researcher can develop understanding of the context to a data set. In the above examples, researchers had prior experience collecting and/or analyzing the data for its original purpose. This gave the researchers access to the original principal investigators for the PBL dataset, the corresponding insight for why certain types of data were collected, and knowledge of what kinds of materials were available beyond the raw data itself. Beyond having prior experience with the data, a researcher can gain contextual understanding for a dataset by examining any supplemental material that goes along with a dataset. If field notes, grant proposals, protocols, or other supporting documents are available, they may provide such context. Additionally, conversation with those who collected the data is also a valuable asset to a secondary researcher trying to understand the context of a dataset, and should be sought when possible.

How well will the dataset help you answer your research question?

After the researcher has developed a contextual understanding of a dataset, they must consider how suitable that data is for investigating their research question. Others have noted that some considerations in establishing suitability are the quality of data and the format of the data[18, 19]. Quality of data refers to how much information the data contains, and how consistent the information is within the dataset. In Example 3 above, the researcher considered the quality of interview data by determining whether or not there were prompts in the protocol that matched his research needs, and then examining responses to those prompts to see if they were applied consistently and with enough detail to be useful.

Data format refers to the kinds of data that will be encountered and how they are stored. Different kinds of data include interviews, observations, or written responses. The importance of data type is discussed within each of the examples above. Different types of storage include text, audio, and video. Researchers should consider, for example, whether they would need transcripts of interviews, and whether those are available. Alternatively, if video would provide necessary information such as in Example 1 above, researchers should consider whether original video is available to them.

What kind of access do you have to the dataset?

Access is an important consideration for a researcher performing qualitative data mining[18]. While some data and results from certain studies are required to be publically available under certain grants, there are likely limitations to the availability of most data, especially data that contains information that can identify study participants. If you are considering a dataset for secondary analysis and are *not* already involved with its collection or analysis, the first step

towards using it is to contact one of the original investigators involved with its collection. Investigators can let you know what conditions the institutional review board (IRB) protocol mandated for that data, whether certain parts of a dataset are available to individuals not connected to that protocol, and whether it would be possible for a new investigator to be added to the IRB protocol to obtain access to the data.

Access is easier to gain if you are previously associated with a dataset. In each of the examples, for instance, the researchers had prior access to the data. Researchers who do not have prior access will likely need to be prepared to answer questions that original investigators might have, such as the purpose of the proposed secondary research, and what kinds of data are of interest. Secondary researchers in this situation should note that original investigators may not be willing to share some data if they are not required to do so, and may also be prohibited from sharing some data due to IRB restrictions.

Are there any other logistical considerations associated with a dataset?

Finally, researchers performing qualitative data mining should be aware of other logistical considerations with using data for secondary analysis. This includes time associated with gaining access to data, time needed to understand the context of the data, and whether the dataset has been overused [18, 19]. As noted in the previous section, there are a number of considerations for gaining access to existing data. Some of these may be time consuming or challenging in their own ways. Additionally, once access has been granted to the data, there may be additional time needed to develop understanding of the context for that dataset, both by talking to the original investigators and by beginning to examine available materials. Lastly, if the data has already been used for multiple secondary analyses, the researcher may want to consider whether the data has been overused.

Where to start looking

We propose some places to start when it comes to qualitative data mining. Depending on your current situation, there are multiple different ways to find existing data sets. If you are a current graduate student, a good place to start is with researchers in your academic department. Learn about past projects by talking to them, or reviewing their publication history. You may also want to ask about data sets they have that are not currently being used.

Outside of institutions you are currently associated with, potential qualitative data miners can also look to conferences and other professional networking opportunities as a way to meet potential collaborators and identify useful datasets. Networking with more individuals can be useful even if you do not have a dataset or research idea that is fully developed: contacts you make now could still help you in the future when you do develop your research agenda.

Finally, there are a few existing resources that can be used for data mining. There are some existing databases of qualitative data such as The UK Data Archive[21]. Additionally, there is an NSF requirement for all NSF grantees to make their data available to other researchers[22]. Some researchers with NSF grants might have online locations for posting the publically available components of their data. Others could be contacted to request availability of certain data.

Potential qualitative data miners could look over publically available lists of NSF grants for information that might be useful to them.

Conclusions

Qualitative data mining can be a useful option for researchers with limited resources. In this paper, we have provided both guidelines for qualitative data mining, and examples demonstrating those guidelines. In addition to this paper, we suggest those who are interested in this type of research consult other publications on secondary analysis[3-6, 18-20].

Acknowledgements

Bibliography

1. Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. 2009, New York, NY: Springer-Verlag New York.
2. Han, J., M. Kamber, and J. Pei, *Data mining: concepts and techniques: concepts and techniques*. 2011, Waltham, MA: Elsevier.
3. Seale, C., *Secondary Analysis of Qualitative Data*, in *Qualitative Research*, D. Silverman, Editor. 2010, SAGE.
4. Glass, G.V., *Primary, secondary, and meta-analysis of research*. Educational researcher, 1976: p. 3-8.
5. Heaton, J., *Secondary analysis of qualitative data: an overview*. Historical Social Research/Historische Sozialforschung, 2008: p. 33-45.
6. Corti, L. and P. Thompson, *Secondary Analysis of Archive Data*, in *Qualitative Research Practice*, C. Seale, et al., Editors. 2004, SAGE: London.
7. Dörre, J., P. Gerstl, and R. Seiffert. *Text mining: finding nuggets in mountains of textual data*. in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. 1999. ACM.
8. Newstetter, W.C., et al. *Does Lecturing Put You to Sleep: How Problem-based Learning Can Awaken Your Classroom*. in *Conference on Higher Education Pedagogy*. 2013. Blacksburg, VA.
9. Hunter, D.A.N., H.M. Matusovich, and M.C. Paretti. *Work in Progress: A Model for Facilitating Problem Based Learning*. in *Frontiers in Education Conference*. 2012. Seattle, WA.

10.   Hunter, D.A.N., Implementing Problem-based Learning in Introductory Engineering Courses: A Qualitative Investigation of Facilitation Strategies, PhD Dissertation in Dept. of Eningeering Education. 2015, Virginia Tech: Blacksburg, VA.
11.   Matusovich, H.M., et al., Faculty Motivation: A Gateway to Transforming Engineering Education. Journal of Engineering Education, 2014. 103(2): p. 302-330.
12.   Brown, P. and H. Matusovich. Unlocking Student Motivation: Development of an Engineering Motivation Survey in Americaon Society of Engineering Education Annual Conference. 2013. Atlanta, GA.
13.   Jones, B.D., et al., Reationships among Students' Perceptions of a First-Year Engineering Design Course and their Engineering Identification, Motivational Beliefs, Course Effort, and Academic Outcomes. International Journal of Engineering Education, 2014. 30(6(A)): p. 1340-1356.
14.   Paretti, M.C., et al., Work in Progress: A Mixed-Methods Study of the Effects of First-Year Project Pedagogies on the Motivation, Retention, and Career Plans of Women in Engineering, in 40th ASEE/IEEE Frontiers in Education Conference. 2010: Washington, DC.
15.   Mccord, R., Thinking About Thinking in Study Groups: Studying Engineering Students' Use of Metacognition in Naturalistic Setting, PhD Dissertation in Dept. of Engineering Education. 2014, Virginia Tech: Blacksburg, VA.
16.   Brown, P., From Content to Choice: Exploring Career Goals in Undergraduate Engineering Students, PhD Dissertation in Department of Engineering Education. Expected May 2016, Virginia Tech: Blacksburg, VA.
17.   Creswell, J.W., *Research design: Qualitative, quantitative, and mixed methods approaches*. 2013: Sage.
18.   Corti, L. and L. Bishop, *Strategies in Teaching Secondary Analysis of Qualitative Data.* Forum: Qualitative Social Research, 2005. **6**(1).
19.   Turner, P.D., *Secondary Analysis of Qualitative Data*. 1997, Presented at the Annual Meeting of the American Educational Research Association, March 24-28, 1997: Chicago, IL.
20.   Hammersley, M., *Can We Re-Use Qualitative Data Via Secondary Analysis? Notes on Some Terminological and Substantive Issues.* Sociological Resaerch Online, 2010. **15**(1).
21.   *The UK Data Archive.* Available from: https://www.ukdataservice.ac.uk/.
21.   *NSF Data Archiving Policy*. Available from: http://www.nsf.gov/sbe/ses/common/archive.jsp.