

Analyzing qualitative data

Prepared by Ashley Suh, June 2021

1

This document examines the following:

- (1) How do you analyze qualitative data?
 - (2) How do you develop a codebook when analyzing qualitative data?
 - (3) How have others coded or analyzed qualitative data in the past?
-

Table of Contents

Thematic analysis	2
Using thematic analysis in psychology	2
Developing a codebook	4
Developing and Using a Codebook for the Analysis of Interview Data: An Example from a Professional Development Research Project	4
Codebook Development for Team-Based Qualitative Analysis	6
How previous work has coded qualitative data	9
Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices	9
Why Authors Don't Visualize Uncertainty	9

Thematic analysis

Using thematic analysis in psychology

Virginia Braun & Victoria Clarke

<https://biotap.utk.edu/wp-content/uploads/2019/10/Using-thematic-analysis-in-psychology-1.pdf.pdf>

- This paper outlines what thematic analysis is and separates the process into 6 phases. The authors explain *how* to perform thematic analysis, and offer tips / guidelines for analyzing themes in qualitative data
- Each theme should capture something important about the data in relation to the research question, and represents some level of *patterned* response or meaning within the data set
- Themes should be prevalent
- Be flexible about themes
- The key-ness of a theme is not dependent on quantifiable measures but rather on whether it captures something important in relation to the overall research question
- There is no right or wrong answer for determining prevalence of a theme
 - It can be the # of different speakers who articulated a theme, or across the entire data set, or each individual occurrence of the theme across the entire data set
- JUST BE CONSISTENT WITH YOUR PROCESS
- Themes should be a *rich description of the data set, or a detailed account of one particular aspect*
- You can identify themes or patterns in one of two primary ways:
 - First, in an inductive, or bottom up way. It means the themes identified are strongly linked to the data themselves. It's similar to grounded theory. If the data have been collected specifically for the research (e.g. interviews or focus groups) the themes identified may bear little relation to the specific questions that were asked of the participants. They would also not be driven by the researcher's theoretical interest in the area or topic. It's a process of coding the data *without* trying to fit it into a pre-existing coding frame, or the researcher's analytic preconceptions. In this sense, this form of analysis is data-driven.
 - Second, in a theoretical or deductive or top down way. This lets researchers drive the analysis by the theoretical or analytic interest in the area. Tends to provide less a rich description of the data overall, and more a detailed analysis of some aspect of the data. You can either code for a quite specific research question (which maps onto the more theoretical approach) or the specific research question can evolve through the coding process (which maps onto the inductive approach).
- There are 6 phases for conducting thematic analysis:

Analyzing qualitative data

Prepared by Ashley Suh, June 2021

Table 1 Phases of thematic analysis

Phase	Description of the process
1. Familiarizing yourself with your data:	Transcribing data (if necessary), reading and re-reading the data, noting down initial ideas.
2. Generating initial codes:	Coding interesting features of the data in a systematic fashion across the entire data set, collating data relevant to each code.
3. Searching for themes:	Collating codes into potential themes, gathering all data relevant to each potential theme.
4. Reviewing themes:	Checking if the themes work in relation to the coded extracts (Level 1) and the entire data set (Level 2), generating a thematic 'map' of the analysis.
5. Defining and naming themes:	Ongoing analysis to refine the specifics of each theme, and the overall story the analysis tells, generating clear definitions and names for each theme.
6. Producing the report:	The final opportunity for analysis. Selection of vivid, compelling extract examples, final analysis of selected extracts, relating back of the analysis to the research question and literature, producing a scholarly report of the analysis.

- After you've transcribed your interviews, start generating initial codes. Each code represents a feature from the raw data that seems interesting. This feature should be the most basic segment, or element, of the raw data such that it can be assessed in a meaningful way regarding the phenomenon we're studying. Example:

Data extract	Coded for
it's too much like hard work I mean how much paper have you got to sign to change a flippin' name no I I mean no I no we we have thought about it ((inaudible)) half heartedly and thought no no I jus- I can't be bothered, it's too much like hard work. (Kate F07a)	1. Talked about with partner 2. Too much hassle to change name

- Advice for initial coding:
 - Code for as many potential themes/patterns as possible (time permitting)
 - Code extracts of data inclusively - i.e., keep a little of the surrounding data if relevant, a common criticism of coding is that the context is lost
 - Remember that you can code individual extracts of data in as many different 'themes' as they fit into - so an extract may be uncoded, coded once, or coded many times, as relevant
 - No data set is without contradiction, so your coding should not have to smooth out or ignore inconsistencies in the data

Analyzing qualitative data

Prepared by Ashley Suh, June 2021

4

Developing a codebook

Developing and Using a Codebook for the Analysis of Interview Data: An Example from a Professional Development Research Project

Jessica T. DeCuir-Gunby, Patricia L. Marshall, and Allison W. McCulloch

<https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.961.3663&rep=rep1&type=pdf>

- This article gives specific steps on how to create a codebook for coding interview data.
- Codebooks provide a formalized operationalization of the codes
 - Coding is the *assigning* of the codes to raw data
- Codes are defined as “tags or labels for assigning units of meaning to the descriptive or inferential information compiled during a study”

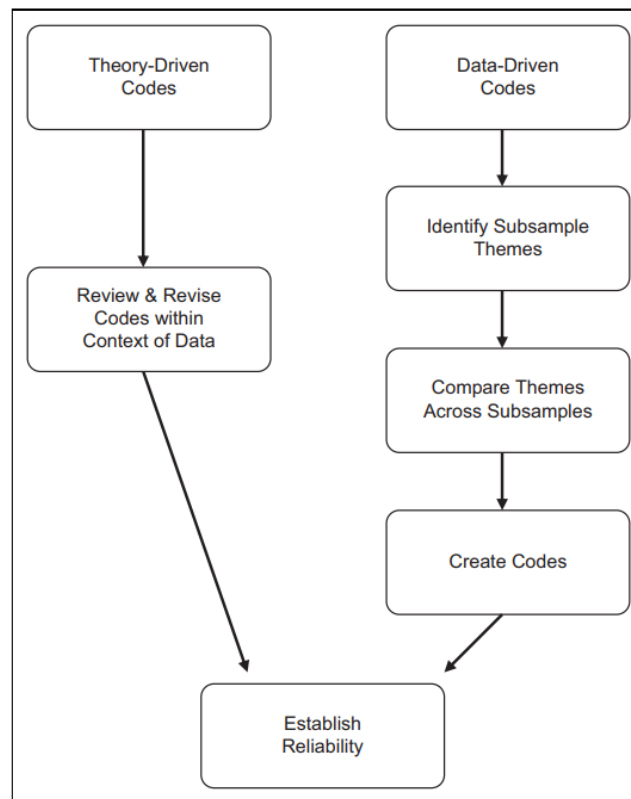


Figure 2. Steps for developing a codebook.

- Developing codes are the first step in analyzing interview data
- Codes should be assigned to chunks of data, usually phrases, sentences, or paragraphs
- Codes are primarily developed from:
 - Existing theories / concepts (theory-driven)
 - Requires constant revisiting of theory
 - Generate the code
 - Review and revise the code in context of the data
 - Determine the reliability of coders and the code

Analyzing qualitative data

Prepared by Ashley Suh, June 2021

5

- Actual raw data (data-driven)
 - Requires repeated examination of the raw data
 - Reduce raw information
 - Identify subsample themes
 - Compare themes across subsamples
 - Create codes
 - Determine reliability of codes
- Codebooks may need to be iterated on, including revising definitions, as researchers gain clearer insights into the interview data
- The more specificity in a codebook, the easier it is for coders to distinguish between codes and determine examples from nonexamples of individual codes
- **CODEBOOKS SHOULD CONSIST OF THREE COMPONENTS:**
 - **Code name / label**
 - **Full definition (an extensive definition that has inclusion and exclusion criteria)**
 - **An example**
- Code labels should be conceptually meaningful, clear and concise, and close to the data (Boyatzis 1998). Therefore, code labels will be iteratively revised.
- Definitions for a code should be specific, but also comprehensive of the constructs you're trying to capture in your research question
 - Can these definitions be easily interpreted by others using your code labels in the future?
 - Make sure your definitions don't contain extraneous information
- Add quotes to definitions to provide good examples of your label
- Example of codes:

Analyzing qualitative data

Prepared by Ashley Suh, June 2021

Table 1. Sample Theory-Driven Codes, Definitions, and Examples

Code	Description	Example
Conception-based reference	Teacher states or alludes to a belief that <i>learners must construct meanings</i> for mathematical ideas <i>on the basis of the learner's existing conceptions</i> that may be quite different from those of the teacher	"Because that then takes—allows the children to use all different strategies to apply and figure out an answer. And there's not that one right answer, which I think is important for me to get kids away from. There's more than one way. It's okay you and I didn't do it the same way."
Cultural referencing	Teacher makes direct/indirect reference to specific elements of students' culture/background (e.g., race, socioeconomic status, language, other outside of school experiences, etc.) that may impact the teaching-learning process	"I think you can see it in a child who—for lack of a better term, is street-wise. You know, they understand what a concept of a number. Like if you have \$5 and you know that you can buy X, Y, and Z with \$5. Then they know five is more than two."
Procedural understanding description	Teacher describes or gives examples of what she believes characterizes procedural understanding.	"So I think procedure is just a rote kind of thing but you don't [know how] it works but you just do it. That's all you know."

- Decide if you code per line, sentence, paragraph, etc. You decide by figuring out which produces the most meaningful themes in the context of the interview data
- Once you develop theory-driven codes, you visit the interviews to figure out what themes were not captured from the theory.
- Code themes per interview
- Once you've developed data-driven codes, the final step is determining the utility/reliability of the codes
 - This can be reported by calculating a basic proportion of agreement from the analyzers of the codes / codebook
 - The number of agreements divided by the total number of agreements + disagreements
 - A reliability of 90% or better is necessary for maximum consistency of coding
- Discuss the examples and non-examples of a code
- Re-examine the data-driven codes in relation to the theory-driven codes to identify and eliminate any overlap

Codebook Development for Team-Based Qualitative Analysis

Kathleen M MacQueen, Eleanor LcLellan, Kelly Kay and Bobby Milstein (CDC)

<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.113.1305&rep=rep1&type=pdf>

Six basic components of a codebook:

1. The code / label

Analyzing qualitative data

Prepared by Ashley Suh, June 2021

7

2. Brief description
3. Full description
4. Guidelines for when to use the code
5. Guidelines for when not to use the code
6. Examples

^ This formalization is condensed to 3 steps in the paper above this one.

- The authors suggest using a relational database to store codes, where the table represents your codebook, the rows represent each code, and the columns represent the attributes of your code.
- When your qualitative data is represented as brief responses to open-ended questions, there is little need to develop structural codes because the data is prestructured by question and participant. For us, our goal is to code the text in a way that the information can be combined meaningfully with a qualitative database. The codes are used primarily to signal the presence or absence of particular pieces of information. The open-ended responses can then be summarized in a 0/1 matrix where the rows are labeled with participant identification numbers and the columns with the codes; the cells are filled with either a 0 (no info present) or a 1 (to indicate that info is present). You can also attach values to the codes. It is suggested to limit the number of codes when analyzing open-ended responses.

TIPS FOR DEVELOPING A CODEBOOK:

1. Assign primary responsibility to one person for creating, updating, and revising a given codebook. This will ensure the codebook is consistent, and it will minimize ambiguities. Any edits to the codebook by other authors should be clarified with the codebook maintainer.
2. Schedule regular meetings where the coding team reviews each code and definition in the codebook. The codebook should be viewed as an analytic tool.
3. Establish a coding process that seeks to enhance intercoder agreement. Coders can reasonably handle 30-40 codes at one time. If using > 40 codes, multiple stages will be required.
4. Develop a written plan for segmenting text at the same time the codebook is being developed. Ashley proposes that we segment the text by each paragraph spoken by the participant. This paragraph may be represented by an individual response to a question from the interview (if the overall concept does not delineate during the individual spoken paragraph), or this paragraph may be represented by the participant stating a new idea (i.e., a new paragraph) when they are giving a response to a question. **For the participants who give 1 or 2 word answers vs. those who give multi-paragraph responses ... it may depend on the participant, may depend on the question -- but as part of coding, it's our job to separate the insight (i.e., the actual response) to any tangential discussion that isn't relevant.**
5. Establish intercoder agreement measures early in the process of codebook development. Agree how we will assess authors' agreement in codes.
6. When defining codes, do not assume anything is obvious; always state specifically what the code should and should not capture. Do not use shorthand references.

Analyzing qualitative data

Prepared by Ashley Suh, June 2021

8

7. Don't clutter the codebook with dead weight: throw out codes that don't work, and rework definitions for codes as problems arise. Be flexible and don't let codes be too specific such that they capture a very specific response from a participant. Using generic codes can capture many responses on a given topic.
8. Accept the fact that text will need to be recoded as the codebook is refined.

How previous work has coded qualitative data

Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices

Sara Alspaugh and Nava Zokaei and Andrea Liu and Cindy Jin and Marti A. Hearst

- [Section 3.3] Three authors collaborated to iteratively develop a code book based on the interview transcripts, resulting in 75 codes in addition to codes for 82 software tools.
- [Codebook]: The codes are organized into a hierarchy with the following top-level categories:
 - Background (e.g., data characteristics, professional role)
 - Workflow stage (e.g., exploration goals, data collection challenges, analysis methods)
 - Tools used (e.g., Jupyter, Tableau), and subcodes (e.g., pro, con)
 - Desired tools and features
 - Homegrown automation (e.g., self-created scripts and tools)
- Coding took place across 5 months and was primarily conducted by two of the authors
- For coding, a participant's utterance was defined as one turn taken in conversation, for a total of 8683 utterances. Each utterance was coded as a whole, and each utterance could be assigned more than one code.
- 3 pass were taken over each transcript:
 - Pass 1: two coders independently applied codes to each utterance in the transcript
 - Pass 2a: one coder reviewed the codes given by her partner to each utterance, and updated her codes with any she felt they had missed or that upon reflection, she agreed with.
 - Pass 2b: the other coder did the same with the first coder's updated codes. This resolved some coding differences that resulted not from a fundamental disagreement but were merely oversights
 - Pass 3: any remaining differences in codes represents true differences of opinion, so in pass 3, the third coder made a tie-breaking pass. Due to the level of care in developing of the code book and coding method, the average agreement between coders prior to the tie-breaking pass was .91 as measured by Cohen's kappa [13]

Why Authors Don't Visualize Uncertainty

Jessica Hullman

- Section 3 doesn't describe the coding process as well, but does have a citation for open coding
- Started with open coding to identify themes, then iterated over the responses several times until the codes stabilized [11]*. For both survey and interview results, she categorized:
 - Understandings of the term "uncertainty visualization"
 - Descriptions of how participants depicted uncertainty (if applicable)
 - Rationales for not expressing uncertainty
 - Also coded statements that seemed aimed at describing a perceived status quo, which became the basis for the rhetorical model
 - Perceptions of the value of uncertainty

[11]: **Qualitative research designs: Selection and implementation**

Analyzing qualitative data

Prepared by Ashley Suh, June 2021

10

<https://biotap.utk.edu/wp-content/uploads/2019/10/qualitative-research-designs.pdf>

- A "**Narrative research**" study involves analyzing your results by trying to "retell" the story into (at times) chronological events lined up across participants from the study.

Enterprise Data Analysis and Visualization: An Interview Study

Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer

- Section 3.3 gives a short explanation of coding, although with no citation
- "Iterative coding method" - grouped common practices, tools, challenges and organizational issues into high level categories
- Iterated and refined those categories as they gathered more data
- They describe the types of analysts encountered in their interviews, and the social context in which those analysts perform analysis
- Then the authors describe recurring patterns in the analytic process and enumerate the most common and severe challenges faced
- Throughout, the authors use representative quotes from respondents to support their analytic claims (i.e., the authors' themes)

Developing the codebook

1. Start open coding, "breaking data apart and delineating concepts to stand for blocks of raw data". Explore different ideas & their meanings within the raw data.
2. Create codes or concepts during open coding.

Analyze the codes through the process of axial coding (higher level of coding that lets us identify any connections that may exist between codes)