Thematic analysis – general notes

Thematic analysis is one of the most common forms of analysis in **qualitative research**. It emphasizes pinpointing, examining, and recording patterns (or "themes") within data. Themes are patterns across data sets that are important to the description of a phenomenon and are associated to a specific research question. Thematic analysis is best thought of as an umbrella term for a variety of different approaches, rather than a singular method. Different versions of thematic analysis are underpinned by different philosophical and conceptual assumptions and are divergent in terms of procedure.

Description

Thematic analysis is used in qualitative research and focuses on examining themes within data. This method emphasizes organization and rich description of the data set. Thematic analysis goes beyond simply counting phrases or words in a text and moves on to identifying implicit and explicit ideas within the data. **Coding** is the primary process for developing themes within the raw data by recognizing important moments in the data and encoding it prior to interpretation. The interpretation of these codes can include comparing theme frequencies, identifying theme co-occurrence, and graphically displaying relationships between different themes. Most researchers consider thematic analysis to be a very useful method in capturing the intricacies of meaning within a data set.

There is a wide range as to what a "data set" entails (see **qualitative data**). Texts can range from a single-word response to an open-ended question or as complex as a body of thousands of pages. As a consequence, data analysis strategies will likely vary according to size. Most qualitative researchers analyze transcribed in-depth interviews that can be 2-hours in length, resulting in nearly 40 pages of transcribed data per respondent. Also, it should be taken into consideration that complexity in a study can vary according to different data types.

Thematic analysis takes the concept of supporting assertions with data from **grounded theory**. This work is designed to construct theories that are grounded in the data themselves. This is reflective in thematic analysis because the process consists of reading transcripts, identifying possible themes, comparing and contrasting themes, and building theoretical models.

Thematic analysis is also related to **phenomenology** in that it focuses on the human experience subjectively. This approach emphasizes the participants' perceptions, feelings and experiences as the paramount object of study. Rooted in humanistic psychology, phenomenology notes giving voice to the "other" as a key component in qualitative research in general. This allows the respondents to discuss the topic in their own words, free of constraints from fixed-response questions found in quantitative studies.

Like most research methods, this process of data analysis can occur in two primary ways— inductively or deductively. In an inductive approach, the themes identified are strongly linked to the data because assumptions are data-driven. This means that the process of coding occurs without trying to fit the data into a pre-existing model or frame. It is important to note that throughout this inductive process, it is not possible for the researchers to free themselves from their theoretical epistemological responsibilities. Deductive approaches, on the other hand, are theory-driven. This form of analysis tends to be less descriptive overall because analysis is limited to the preconceived frames. The result tends to focus on one or two specific aspects of the data that were determined prior to data analysis. The choice between these two approaches generally depends on the researchers' epistemologies (see **epistemology**).

Theme

A theme represents a level of patterned response or meaning from the data that is related to the research questions at hand. Determining what can be considered a theme can be used with

deciding prevalence. This does not necessarily mean the frequency at which a theme occurs, but in terms of space within each data item and across the data set. It is ideal that the theme will occur numerous times across the data set, but a higher frequency does not necessarily mean that the theme is more important to understanding the data. A researcher's judgement is the key tool in determining which themes are more crucial. A potential data analysis pitfall occurs when researchers use the research question to code instead of creating codes and fail to provide adequate examples from the data. Eventually, themes need to provide an accurate understanding of the "big picture"

There are also different levels at which themes can be identified—semantic and latent. A thematic analysis generally focuses wholly or mostly on one level. Semantic themes attempt to identify the explicit and surface meanings of the data. The researcher does not look beyond what the participant said or wrote. In this instance, the researcher wishes to give the reader a sense of the important themes. Thus, some depth and complexity is lost. However, a rich description of the *entire* data set is represented. Conversely, latent themes identify *underlying* ideas, patterns, and assumptions. This requires much interpretation of the data, so researchers might focus on one specific question or area of interest across the majority of the data set.

A theme is different from a code. Several texts recommend that researchers "code for themes". This can be misleading because the theme is considered the outcome or result of coding, not that which is coded. The code is the label that is given to particular pieces of the data that contribute to a theme. For example, "SECURITY can be a code, but A FALSE SENSE OF SECURITY can be a theme."

Methodological issues

Reflexivity journals

Given that qualitative work is inherently interpretive research, the biases, values, and judgments of the researchers need to be explicitly acknowledged so they are taken into account in data presentation. This type of openness is considered to be positive in the qualitative community. Researchers shape the work that they do and work as the instrument for collecting and analyzing data. In order to acknowledge the researcher as the tool of analysis, it is necessary for one to create and maintain a reflexivity journal.

The reflexivity process can be described as the documenting close reflections of potential findings and implications of the research study. Reflexivity journals are often referred to as analytic memos or memo writing, which can be useful for reflecting on emergent patterns, themes and concepts. Throughout the coding process researchers should have detailed records of the development of each of their codes and potential themes. In addition, changes made to themes and connections between themes are incorporated into the final report to assist the reader in understanding decisions that were made throughout the coding process.

Once fieldwork and interviews are complete and researchers are beginning the data analysis stages, they should take notes from the transcription and interviews. Researchers can take notes by writing down any words that may be of use during data analysis in a journal or notebook. The logging of ideas for future analysis can aid in getting thoughts and reflections written down and may serve as a reference for potential coding ideas as one progresses from one stage to the next in the thematic analysis process. Items written in journal do not have to be accurate or final but instead should contain considerations for further analysis. Researchers must take into consideration that analytic memos will assist them in the future coding of potential overreaching themes.

While working on reflexivity journal entries it is important to make certain that notes written in journals are different from the data. The use of italics, bolding words, and adding brackets will assist in showing distinctions between data and journaling. Researchers should write their reflexivity notes fully avoiding abbreviations. This will assist the researcher in the final stages of analysis and through the process of data complication and reduction. Auerbach & Silverstein (2003) suggest keeping a log of concerns with the research, theoretical framework, central

research questions, goals, and major issues to help focus on the coding process. Analytic memos reveal information about the researchers thinking process pertaining to the codes and categories that have emerged throughout the analysis process. One of the most critical outcomes of qualitative data analysis is to interpret how each individual components of the study relate to each other, in particular researchers should focus on observations of the population to gain an image of the bigger picture that may lead to universal observations. Emerson, Fretz & Shaw (1995) recommend the following questions should be considered when coding fieldwork notes:

Coding practice

Questions to consider whilst coding may include:

- What are people doing? What are they trying to accomplish?
- How exactly do they do this? What specific means or strategies are used?
- How do members talk about and understand what is going on?
- What assumptions are they making?
- What do I see going on here? What did I learn from note taking?
- Why did I include them?

Such questions are generally asked throughout all cycles of the coding process and the data analysis, while notable themes are noted. A reflexivity journal is often used to identify potential codes that were not initially pertinent to the study.

Sample size considerations

There is little reliable guidance on what sample size is needed for a thematic analysis, with suggestions ranging from 6 to 400+ depending on the type of data collection and size of the project.

It is common not to specify the number needed at the outset, but for decisions to be made as the research proceeds. One approach is to continue to include material (*e.g.* further interviews) until no further themes are found; that is, until saturation. The number needed to reach saturation has been investigated empirically, but such approaches do not readily allow a prospective estimation of the sample needed. Fugard & Potts (2015) offered a prospective, quantitative tool to support thinking on sample size by analogy to quantitative sample size estimation methods. Lowe et al. (2018) propose quantitative, probabilistic measures of degree of saturation that can be calculated from an initial sample and used to estimate the sample size required to achieve a specified level of saturation. Their analysis indicates that commonly-used binomial sample size estimation methods may significantly underestimate the sample size required for saturation.

Phases

Phase	Process	Result	Reflexivity Journal Entries[
Phase 1	Read and re-read data in order to become familiar with what the data entails, paying specific attention to patterns that occur.	Preliminary "start" codes and detailed notes.	List start codes in journal, along with a description of what each code means and the source of the code.
Phase 2	Generate the initial codes by documenting where and how patterns occur. This happens through data reduction where the researcher collapses data	Comprehensive codes of how data answers research question.	Provide detailed information as to how and why codes were combined, what questions the researcher is

	into labels in order to create categories for more efficient analysis. Data complication is also completed here. This involves the researcher making inferences about what the codes mean.		asking of the data, and how codes are related.
Phase 3	Combine codes into overarching themes that accurately depict the data. It is important in developing themes that the researcher describes exactly what the themes mean, even if the theme does not seem to "fit". The researcher should also describe what is missing from the analysis.	List of candidate themes for further analysis.	Reflexivity journals need to note how the codes were interpreted and combined to form themes.
Phase 4	In this stage, the researcher looks at how the themes support the data and the overarching theoretical perspective. If the analysis seems incomplete, the researcher needs to go back and find what is missing.	Coherent recognition of how themes are patterned to tell an accurate story about the data.	Notes need to include the process of understanding themes and how they fit together with the given codes. Answers to the research questions and data-driven questions need to be abundantly complex and well-supported by the data.
Phase 5	The researcher needs to define what each theme is, which aspects of data are being captured, and what is interesting about the themes.	A comprehensive analysis of what the themes contribute to understanding the data.	The researcher should describe each theme within a few sentences.
Phase 6	When the researchers write the report, they must decide which themes make meaningful contributions to understanding what is going on within the data. Researchers should also conduct "member checking". This is where the researchers go back to the sample at hand to see if their description is an accurate representation.	A thick description of the results.	Note why particular themes are more useful at making contributions and understanding what is going on within the data set. Describe the process of choosing the way in which the results would be reported.

Phase 1: Becoming familiar with the data

The initial phase in thematic analysis is for researchers to familiarize themselves with the data. Prior to reading the interview transcripts, researchers may create a "start list" of potential codes. These start codes should be included in a reflexivity journal with a description of representations of each code and where the code is established. Analyzing data in an active way will assist researchers in searching for meanings and patterns in the data set. At this stage, it is tempting to skip over the data; however, this will aid researchers in identifying possible themes and patterns. Reading and re-reading the material until the researcher is comfortable is crucial to the initial phase of analysis. While becoming familiar with the material, note-taking is a crucial part of this step in order begin developing potential codes.

Transcription

After completing data collection, the researcher needs to begin transcribing the data into written form. For further information on this process, please refer to transcription. Transcription of the data is imperative to the dependability of analysis. Transcribed data can come from television programs, interviews (see interviewing), and speeches, among others.

 Criteria for transcription of data must be established before the transcription phase is initiated to ensure that dependability is high Inconsistencies in transcription can produce biases in data analysis that will be difficult to identify later in the analysis process. The protocol for transcription should explicitly state criteria of transcription. Inserting comments like "*voice lowered*" will signal a change in the speech. In this stage, it is especially important to draw upon non-verbal utterances and verbal discussions to lead to a richer understanding of the meaning of data. A general guideline to follow when transcribing includes a ratio of 15 minutes of transcription for every 5 minutes of dialog.

After this stage, the researcher should feel familiar with the content of the data and should be able to identify overt patterns or repeating issues in one or more interviews. These patterns should be recorded in a reflexivity journal where they will be of use when coding and checking for accuracy. Following the completion of the transcription process the researcher's most important task is to begin to gain control over the data. At this point, it is important to mark data that addresses the research question. This is the beginning of the coding process.

Phase 2: Generating initial codes

The second step in the thematic analysis is generating an initial list of items from the data set that have a reoccurring pattern. This systematic way of organizing, and gaining meaningful parts of data as it relates to the research question is called coding. The coding process evolves through an inductive analysis and is not considered to be a linear process, but a cyclical process in which codes emerge throughout the research process. This cyclical process involves going back and forth between phases of data analysis as needed until you are satisfied with the final themes. Researchers conducting thematic analysis should attempt to go beyond surface meanings of the data to make sense of the data and tell an accurate story of what the data means.

The coding process is rarely completed the first time. Each time, researchers should strive to refine codes by adding, subtracting, combining or splitting potential codes. Start codes are produced through terminology used by participants during the interview and can be used as a reference point of their experiences during the interview. Dependability increases when the researcher uses concrete codes that are based on dialogue and are descriptive in nature. These codes will facilitate the researcher's ability to locate pieces of data later in the process and identify why they included them. Initial coding sets the stage for detailed analysis later by allowing the researcher to reorganize the data according to the ideas that have been obtained throughout the process. Reflexivity journal entries for new codes serve as a reference point to the participant and their data section, reminding the researcher to understand why and where they will include these start codes in the final analysis. Throughout the coding process, full and equal attention needs to be paid to each data item because it will help in the identification of unnoticed repeated patterns. Coding for as many themes as possible and coding individual aspects of the data may seem irrelevant but can potentially be crucial later in the analysis process. Coding also involves the process of data

reduction and complication. Reduction of codes is initiated by assigning tags or labels to the data set based on the research question(s). In this stage, condensing large data sets into smaller units permits further analysis of the data by creating useful categories. In-vivo codes are also produced by applying references and terminology from the participants in their interviews. Coding aids in development, transformation and re-conceptualization of the data and helps to find more possibilities for analysis. Researchers should ask questions related to the data and generate theories from the data, extending past what has been previously reported in previous research.

Data reduction

Coding can be thought of as a means of reduction of data or data simplification. Using simple but broad analytic codes it is possible to reduce the data to a more manageable feat. In this stage of data analysis the analyst must focus on the identification of a more simple way of organizing data. using data reductionism researchers should include a process of indexing the data texts which could include: field notes, interview transcripts, or other documents. Data at this stage are reduced to classes or categories in which the researcher is able to identify segments of the data that share a common category or code. Siedel and Kelle (1995) suggest three ways to aid with the process of data reduction and coding: (a) noticing relevant phenomena, (b) collecting examples of the phenomena, and (c) analyzing phenomena to find similarities, differences, patterns and overlying structures. This aspect of data collection is important because during this stage researchers should be attaching codes to the data to allow the researcher to think about the data in different ways. Coding cannot be viewed as strictly data reduction, data complication can be used as a way to open up the data to examine further. The below section addresses the process of data complication and its significance to data analysis in qualitative analysis.

Data complication

The process of creating codes can be described as both data reduction and data complication. Data complication can be described as going beyond the data and asking questions about the data to generate frameworks and theories. The complication of data is used to expand on data to create new questions and interpretation of the data. Researchers should make certain that the coding process does not lose more information than is gained. Tesch (1990) defines data complication as the process of re-conceptualizing the data giving new contexts for the data segments. Data complication serves as a means of providing new contexts for the way data is viewed and analyzed.

Coding is a process of breaking data up through analytical ways and in order to produce questions about the data, providing temporary answers about relationships within and among the data. Decontextualizing and re-contextualizing help to reduce and expand the data in new ways with new theories.

Phase 3: Searching for themes

Searching for themes and considering what works and what does not work within themes enables the researcher to begin the analysis of potential codes. In this phase, it is important to begin by examining how codes combine to form over-reaching themes in the data. At this point, researchers have a list of themes and begin to focus on broader patterns in the data, combining coded data with proposed themes. Researchers also begin considering how relationships are formed between codes and themes and between different levels of existing themes. It may be helpful to use visual models to sort codes into the potential themes.

Themes differ from codes in that themes are phrases or sentences that identifies what the data *means*. They describe an outcome of coding for analytic reflection. Themes consist of ideas and descriptions within a culture that can be used to explain causal events, statements, and morals derived from the participants' stories. In subsequent phases, it is important to narrow down the potential themes to provide an overreaching theme. Thematic analysis allows for categories or themes to emerge from the data like the following: repeating ideas; indigenous terms, metaphors and analogies; shifts in topic; and similarities and differences of participants' linguistic expression. It is important at this point to address not only what is present in data, but also what is missing from the data. Conclusion of this phase should yield many candidate themes collected throughout the

data process. It is crucial to avoid discarding themes even if they are initially insignificant as they may be important themes later in the analysis process.

Phase 4: Reviewing themes

This phase requires the researchers to search for data that supports or refutes the proposed theory. This allows for further expansion on and revision of themes as they develop. At this point, researchers should have a set of potential themes, as this phase is where the reworking of initial themes takes place. Some existing themes may collapse into each other, other themes may need to be condensed into smaller units.

Specifically, this phase involves two levels of refining and reviewing themes. Connections between overlapping themes may serve as important sources of information and can alert researchers to the possibility of new patterns and issues in the data. Deviations from coded material can notify the researcher that a code may not actually exist. Both of this acknowledgements should be noted in the researcher's reflexivity journal, also including the absence of themes. Codes serve as a way to relate data to a person's conception of that concept. At this point, the researcher should focus on interesting aspects of the codes and why they fit together.

Level 1

Reviewing coded data extracts allows researchers to identify if themes form coherent patterns. If this is the case, researchers should move onto Level 2. If themes do not form clear patterns, consideration of the potentially problematic themes should be considered in addition to determining if data does not fit into the theme. If themes are problematic, it is important to rework the theme and during the process, identification of new themes may emerge. For example, it is problematic when themes do not appear to work or a significant amount of overlap between themes exists. This can result in a weak or unconvincing analysis of the data. If this occurs, data may need to be recognized in order to create cohesive, mutually exclusive themes.

Level 2

Considering the validity of individual themes and how they connect to the data set is crucial to completing this stage. It is imperative to assess whether the potential thematic map accurately reflects the meanings in the data set in order to provide an accurate representation of participants' experiences. Once again, at this stage it is important to read and re-read the data to determine if current themes relate back to the data set. To assist you in this process it is imperative to code any additional items within the themes may have been missed earlier in the initial coding stage. If the potential map works then the researcher should progress to the next phase of analysis. If the map does not work it is crucial to return to the data in order to continue to review and refine existing codes. Mismatches between data and analytic claims reduce the amount of support that can be provided by the data. This can be avoided if the researcher is certain that their interpretations of the data and analytic analysis correspond. Researchers repeat this process until they are satisfied with the thematic map. By the end of this phase, researchers have an idea of what themes are and how they fit together so that they convey a story about the data set.

Phase 5: Defining and naming themes

Defining and refining existing themes that will be presented in the final analysis assists the researcher in analyzing the data within each theme. At this phase, identification of the themes' essences relate to how each specific theme affects the entire picture of the data. Analysis at this stage is characterized by identifying which aspects of data are being captured, what is interesting about the themes, and why themes are interesting.

In order to identify whether current themes contain sub-themes and to discover further depth of themes, it is important to consider themes within the whole picture and also as autonomous themes. Researchers must then conduct and write a detailed analysis to identify the story of each theme and its significance. By the end of this phase, researchers can (1) define what current themes consist of, and (2) explain each theme in a few sentences. It is important to note that researchers begin thinking about names for themes that will give the reader a full sense of the theme and its

importance. Failure to fully analyze the data occurs when researchers do not use the data to support their analysis beyond the content. Researchers conducting thematic analysis should attempt to go beyond surface meanings of the data to make sense of the data and tell an accurate story of what the data means.

Phase 6: Producing the report

After final themes have been reviewed, researchers begin the process of writing the final report. While writing the final report, researchers should decide on themes that make meaningful contributions to answering research questions which should be refined later as final themes. Researchers present the dialogue connected with each theme in support of increasing dependability through a thick description of the result. The goal of this phase is to write the thematic analysis to convey the complicated story of the data in a manner that convinces the reader of the validity and merit of your analysis. A clear, concise, and straightforward logical account of the story across and with themes is important for readers to understand the final report. The write up of the report should contain enough evidence that themes within the data are relevant to the data set. Extracts should be included in the narrative to capture the full meaning of the points in analysis. The argument should be in support of the research question. The final step in producing the report is to include member checking as a means to establish credibility, researchers should consider taking final themes and supporting dialog to participants to elicit feedback.

Advantages and disadvantage

Researchers conducting qualitative analysis should select the most appropriate method to the research question. The method of analysis should be driven by both theoretical assumptions and the research questions. Thematic analysis provides a flexible method of data analysis and allows for researchers with various methodological backgrounds to engage in this type of analysis. Reliability with this method is a concern because of the wide variety of interpretations that arise from the themes, as well as applying themes to large amounts of text. Increasing reliability may occur if multiple researchers are coding simultaneously, which is possible with this form of analysis. To increase reliability with this method researchers should plan for monitoring themes and codes tables throughout the process. This method of analysis contains several advantages and disadvantages, it is up to the researchers to decide if this method of analysis best explains their results.

Advantages

- Flexibility it allows researchers, in that multiple theories can be applied to this process across a variety of epistemologies.
- Well suited to large data sets.
- Allows researchers to expand range of study past individual experiences.
- Great for multiple researchers.
- Interpretation of themes supported by data.
- Applicable to research questions that go beyond an individual's experience.
- Allows for categories to emerge from data.

Disadvantages

Reliability is a concern due to wide variety of interpretations from multiple researchers.

- Thematic analysis may miss nuanced data.
- Flexibility makes it difficult to concentrate on what aspect of the data to focus on.
- Discovery and verification of themes and codes mesh together
- Limited interpretive power if analysis excludes theoretical framework
- Difficult to maintain sense of continuity of data in individual accounts
- Does not allow researchers to make claims about language usage