

Making Thematic Analysis Systematic: The Seven Deadly Sins

David C. Coker

Adjunct Professor, Advanced Education Program

Fort Hays State University Virtual College, United States

Received: July 13, 2021

Accepted: August 30, 2021

Published: August 30, 2021

doi:10.5296/jse.v11i3.18882

URL: <https://doi.org/10.5296/jse.v11i3.18882>

Abstract

Thematic analysis is a methodology with wide use in content analysis and field work through interviews, observations, and focus groups. Despite popularity with researchers, there are several questions and problems with the methodology. There was an absence in exploration and explanation of the methods and hidden steps in thematic analysis. Using a case study with a thematic analysis of the methods of thematic analysis and an autoethnography, issues and concerns were examined across a broad sample of articles published using thematic analysis. Seven major problems permeate thematic analysis research, and recommendations to improve thematic analysis and all qualitative research are presented.

Keywords: thematic analysis, qualitative research, autoethnography, reliability and validity

Introduction

Thematic analysis is a qualitative technique which gained immense popularity following Fereday & Muir-Cochrane (2006) and Braun & Clarke's (2006) treatments. There is versatility in application and construction of thematic analysis, from field work to document analysis. From the development of a codebook and a frequentist approach (Fereday & Muir-Cochrane, 2006) to reflexive thematic analysis (Braun & Clarke, 2006), researchers have wide latitude in theory and construction of methods within the broader methodology.

For all typologies of thematic analysis, defining and enunciating themes are the central goal (Attride-Stirling, 2001). There is the belief data can be examined, reduced, and elaborated with common ideas and structures throughout, with the possibility of producing a taxonomy and hierarchy of features. Moving from the semantic level to the latent level is a primary interpretive concern of most thematic analysis, with the result connections and concepts which were obscured or unclear being developed (Thomas & Harden, 2008). Thematic analysis, like most qualitative methodologies, is both reductionistic and expansionistic.

The hows and whats remain ill-defined for novice and seasoned researchers in much of the existing literature. "If readers are not clear about how researchers analyzed their data or what assumptions informed their analysis, evaluating the trustworthiness of the research process is difficult" (Nowell et al., 2017, para. 2). Lucas and D'Enbeau (2013) found four major problems in qualitative research: moving quickly over analysis, product over the process of analysis, lack of explanation of techniques of analysis, and lack of synthesis. As a researcher who publishes and peer reviews regularly, much of what passes for thematic analysis lacks an explication of a systematic methodology. Simply saying one followed Braun & Clarke (2006) or Fereday & Muir-Cochrane (2006) is not enough. To improve thematic analysis, an instrumental case study by way of autoethnography gives direction on how to alleviate common problems. A thematic analysis of peer reviewed articles adds to the insights gained through the author's praxis.

Methodology

The study used an instrumental case study, with two components. An autoethnography of the researcher's experience in thematic analysis and peer review examined the research processes. A second method was a thematic analysis of peer reviewed articles utilizing thematic analysis. Both methods were employed concomitantly and throughout the analysis.

Yin (2017) stated a case study should be methodologically rigorous, flow from the research questions, and work to confirm or reject rival hypothesis. The how and why comprise two central questions, and the case study richly considered all available evidence and prior theory (Yazan, 2015). Two questions underpinned the case study: How are the unwritten rules, the gray areas, of thematic analysis practiced? What practices could improve thematic analysis? The gray area is how data are encoded and themed to present a holistic view of thematic analysis.

To investigate thematic analysis to describe the nature of the field and recommendations for improvement, the author relies on theory, his adoption of the thematic analysis process, and

his peer review of thematic analysis and qualitative studies. The instrumental case study considers one topic and was both exploratory and explanatory. The gap in the literature was really an absence: Researchers do not discuss and describe the hidden rules and steps which are used to implement the different processes.

An autoethnography and thematic analysis were combined to examine thematic analysis. A reflexive autoethnography details the experiences of a researcher and give a layered approach to provide the readers with a chance to experience and understand the phenomena (Ellis et al., 2011). Three steps define an autoethnography: the author is a full member of the research group, has published in the field, and seeks to expand theoretical underpinnings (Anderson, 2006). The data collection involved the review of publishing seven articles and conducting peer review of 16 articles within a year's time. All data are data, so all experiences with the research process which informed my praxis were included.

All articles written or reviewed for the autoethnography had the rough draft, a note file (most articles were 50-100 pages of notes each), the Excel thematic analysis (no data were ever destroyed or deleted, and copying data to move through different steps allowed a serial comparison of progression), and extensive memoing and thinking by noting to myself in email at all times of days and from any place. Using a detailed record to complement my memory and feelings, notes and outlines were produced for the current study. An iterative fashion, with a reexamination and analysis of the field notes, coupled with the thematic analysis and literature review, broadened the view of the autoethnography to consider both the insider and outsider perspective.

A descriptive thematic analysis was conducted which was included in the case study. There were 35 research, peer reviewed articles selected from 1999-2020. Articles were chosen by searching in Google Scholar. The *a priori* sample size was >30; while not exhaustive, the size was akin to the large sample size needed at minimum in quantitative analysis. The studies could be broken down further: 15 focus groups or focus groups/interviews, 9 content analyses, 7 literature reviews, and 4 questionnaires. The different fields were mainly health and psychology, with few in the other social sciences. Combined with the author's review of thematic analysis for both publishing and peer review, the sample was deemed adequate to explore deductively four major points: sample size and selection, coding methods, thematic methods, and the structure of limitations.

All data were downloaded and an Excel file was set up. The first step was reading and annotating the articles. The second step involved coding the data by brief quotes, descriptions, and categories, as well as memos, notes, and identification of "aha!" moments. The third step happened concurrently involved sorting data by examining repetition, opposites, absences, and the degree to which different codes approximated each other through causes, context, and consequences. Themes were considered by elements, dimensions, and substantive themes and/or theories. Extensive memoing, using email and my cellphone, allowed me to consider and reconsider the connections within the data analysis. The rules commonly used for analogies were considered. After the second article was reviewed, there was a formal rendering of comparing and contrasting to construct themes, and this happened with each

successive article. The fourth step involved checking for the logic and coherence of the themes, as well as checking validity and reliability. Having themes rooted and grounded, parity checks and reconciliation were used to confirm plausibility. Each data source was given a geocode to locate the origins, which allowed the researcher to check each line back to the source in an iterative fashion.

The results of the thematic analysis were merged with the autoethnography to produce a more comprehensive case study. Through a narrative, seven topics originated from the autoethnography, and the thematic analysis both complemented and challenged the introspection. After presenting the results and a discussion, limitations were examined and presented.

Results: The 7 Deadly Sins

The autoethnography and thematic analysis yielded several important characteristics.

Significant overlap existed, but most qualitative research concepts, from the components of triangulation to trustworthiness, build off each concept. Each step supports and complements every other step, but the themes exist alone and add to the understanding of qualitative research and thematic analysis.

Seven deadly sins are discussed and elaborated. Taken together, the seven have the possibility of turning from sins to the virtues embedded in each. Qualitative research has come a long way, but there are significant hurdles and shortcomings which impact the value.

#1 *Justificatio sola fide*. Faith alone.

Reviewing articles in thematic analysis—and many other qualitative paradigms—revealed either the lack of a research paradigm or a hidden one. Either way, reviewers and researchers were expected to use faith alone to have a leap that the methodologies employed were rigorous, systematic, and trustworthy. Four major issues must be decided in a design: paradigm, sample, methods, and the research questions (RQs). *Thou shalt have a design.*

The leap of faith was acute when writing up my first thematic analysis in a research using mission statements for the public schools. Most examples stated they used “Braun and Clarke (2006),” but reading between the lines revealed a system so loose, one could not see a system. The problem was so common, I wrote on peer review after peer review: “You stated you used a system, so show the system.” The Braun/Clarke (2006) use might as well have stated one used technology, like NVivo or Atlas. Either way, one had no idea what was the paradigm or the approach.

I used grounded theory extensively in the mission statement paper, so I thought I could simply transfer my prior knowledge to the six steps espoused by Braun and Clarke. Wow was all I could say. Within the thematic analysis of the 35 articles, the ones who claimed to use grounded theory—doubtful because one never saw or heard theoretical sensitivity—were the clearest on explicating the design. Grounded theory is, to a limited extent, formulaic, but I found using in vivo, descriptive, open codes, focused, and axial did not transfer well to my study. There were a number of reasons why.

My methods were centered around inductive, but there was a focus on actors, actions, and consequences. While I did all the steps of grounded theory, there was a massive redundancy in the content analysis of mission statements. There was eventually over 10,000 units of data! How to write the actors by in vivo, descriptive, etc., proved to be an extreme waste of time. One could not easily write the actors split six different ways, so there had to be change.

Braun and Clarke (n.d.) spoke about a research paradigm by way of inductive, deductive, semantic, latent, and other ways. They might have started the conversation, but one sees the paradigms listed are often orthogonal, omit hybridization, and neglect the scarcity from the value of semantic versus latent. One could be inductive-semantic or deductive-latent, but even these divisions do not prove useful. Peer reviewing activities found researchers frequently omitted the perspective, resulting in the lack of a visible design. Many researchers did not want to disclose the specifics.

The RQs should, to the casual observer, come first in the process. We like to think RQs guide everything. In reality, politics does. The confluence of the paradigm, methods, and sample moderate and mediate the RQs. When one finds the sample or methods cannot answer the question one wished, one adapts to the data. There is always a back and forth between the possible and the impossible.

Visible. Designs should be visible, so the reader is assured a systematic approach researched said phenomena. Research should have an outline: perspective, aim, research question(s), sample, data collection, data analysis, and results. A paradigm influences the perspective, and when one lacks explicit methods, sampling design, and RQs, evaluation of a research project is elusive.

The thematic analysis complemented peer reviews: Either designs were lacking or poorly articulated. Of the 35 thematic articles reviewed, 7 had no design beyond the generic “initial coding” or no mention. What or how initial coding is done never materializes. Next up was inductive and critical realist at 5. Rounding up the list was deductive at 4, and a small number claiming hybrid, grounded, and horizontalism. The orthogonal structure made the claimed paradigms unclear; for example, being a critical realist (or constructivist or any other paradigm) does not clear up whether one is inductive, deductive, or hybrid.

The lack of an explicit design means one cannot evaluate reproducibility and replicability. Perhaps the most labor-intensive sections—the paradigms and coding schema—are missing from many articles. Most thematic research is interpretivist, so semantic and latent discussions have much less weight than, *de minimus*, deductive/inductive/hybrid and the coding mechanism. Another concern, sometimes noted, is failure to account for primary versus secondary qualitative analysis.

I felt deductive research was far easier. There were categories, or the essence of the code. There was a built-in coherence. One lacked the uncertainty and inherent messiness of inductive thematic analysis. With deductive research, one has coding categories and a readily identified BOLO, or be on the look out. What was important was already noted. Peer reviewing deductive articles almost always has a framework. The mission statement article took me in many

different directions and required line-by-line analysis, while the deductive scoping review had the themes directed ahead of time. There was much less antagonizing in the scoping review, as the degrees of each category easily fit into the themes.

Sample. Sample is so important, it needed considered separately. Most researchers, at least initially, probably followed my justification for samples. Whether reading Creswell (2002) or some other popular author, one lifted without thought their recommendations. Reviewers and authors want a reason, so any reason would do. This positivistic outlook in sampling is very troubling for a postpositivist world.

I have seen all kinds of sampling sizes and mechanisms. There are two main axes: person-centered and document analysis. Document analysis typically follows a method which dictates the size; for example, my scoping review included all articles which met my criteria. The mission statement thematic analysis had 80 mission statements. There were rules about selecting high and low schools, but would 100 or 2,0000, etc., have produced more accurate research?

Yes and no. If there is great homogeneity, then one will probably identify most all themes in a sample of 80. In a larger sample, there would be many more fringe statements, and some themes would include different elements and dimensions. Then where or how to stop, as the research base is not always clear for document analysis? My answer was then and now, quantitative research can serve as a guidepost. I picked 80 because it was >30 , and the law of big numbers suggested there would be a greater likelihood of a normal distribution. Also, in multiple regression with three predictors, 80 generally produces adequate power. Alas, what this means is for document analysis, one either follows a rule dictated by the research design (such as including all relevant studies in a meta-analysis) or makes a choice which one can defend as adequate.

Interviews and focus groups have guideposts, too. Guest et al. (2006) make a persuasive claim of 6-12, and many have convincingly shown 4-10 could all provide relevant answers. I wish I could say I saw such sound judgments in sampling. Often there is no reason, and some samples are unreasonably small (such as 9 in one focus group for the entire sample) or extremely lacking in uniformity, such as a small interview group ranging from childhood to an octogenarian. Worse, there are often few details about the sample; a very complete demographic and qualitative description of a sample improves one's review of reliability and validity, as transferability and generalizability cannot be competently evaluated by a reader with a sample largely hidden from view.

Troubling for qualitative researchers is the idea sample sizes are mathematical equations. Guest et al. (2020) present neat analysis how to think of sample size as fitting a formula, something others (e.g., Malterud et al., 2016) decry. In my thematic analysis, 14 articles used interviews or focus groups, and sample sizes ranged from 4-118. I am in the latter group which think sample sizes cannot be reduced to a formula, and there are some sound reasons.

Sample size must be considered holistically through the lens of the entire design. Which would produce more accurate results: a sample of 4 or a sample of 32? There is no way to know by

that information alone. Running four focus groups of one hour each with 8 participants in each would provide *different* information than four one-on-one, in-depth interviews over 8 hours with each. If one has the time, personnel, and resources, then the general rule of thumb “bigger is better” mostly holds true. The one constant is the research design means even the same question can be answered differently depending on sampling and data collection.

#2 *Explicatio temere*. Random analysis.

When I sat down to analyze the data, I naively thought I could simply cut and paste the grounded theory framework to thematic analysis. I had made notes and annotations of mission statements before I started, and I had a framework and direction before starting. There were a myriad of problems. First, though the research was inductive, I was specifically looking at actors, actions, and consequences. Secondly, the coding used in grounded theory was extremely repetitive and unnecessary. For example, in documenting consequences in mission statements, initially 3,924 pieces of information were generated (in vivo, descriptive, open coding, and focused, as well as word counts, location of original information, and the use of a number scheme to track different consequences; if one included memos, constant comparison notes, and aha! moments, the 3,924 would be higher by over one hundred). I quickly felt overwhelmed and unsure how to proceed. *Thou shalt have a coding paradigm.*

I reformed my procedures soon after. First, descriptive codes were used with focused coding. How to track information was a major problem, so I added the idea of a geocode. Simply by affixing a number to each line, sorting meant everything could be rooted. Unlike grounded, rooted meant the origin was pinned and followed all codes generated. Then I sorted by analogy rules, such as similarities, differences, absences, and part of whole, etc. Being rooted, I was able to develop themes and check for accuracy.

There were so many codes, I had to ensure I represented faithfully and accurately what thematic analysis stated. To do this, I sought reconciliation. Generating numbers for each themes meant each linear code could be compared to the original and between each other. It was easy to see a different theme in one idea at different times. I decided to conduct *reconciliation*, or a check that similar descriptive codes resulted in the same theme, so I recoded each one a second time to see if the in vivo, descriptive, and categorical code matched the original.

By the time I conducted a second thematic analysis, I left grounded theory on the shelf and had a cogent system worked out, as shown in Figure 1. I intentionally removed all jargon, and the process emphasizes one does not proceed from step to step. As one step is partially constructed, one proceeds to the next step. There is a back and forth through all the stages.

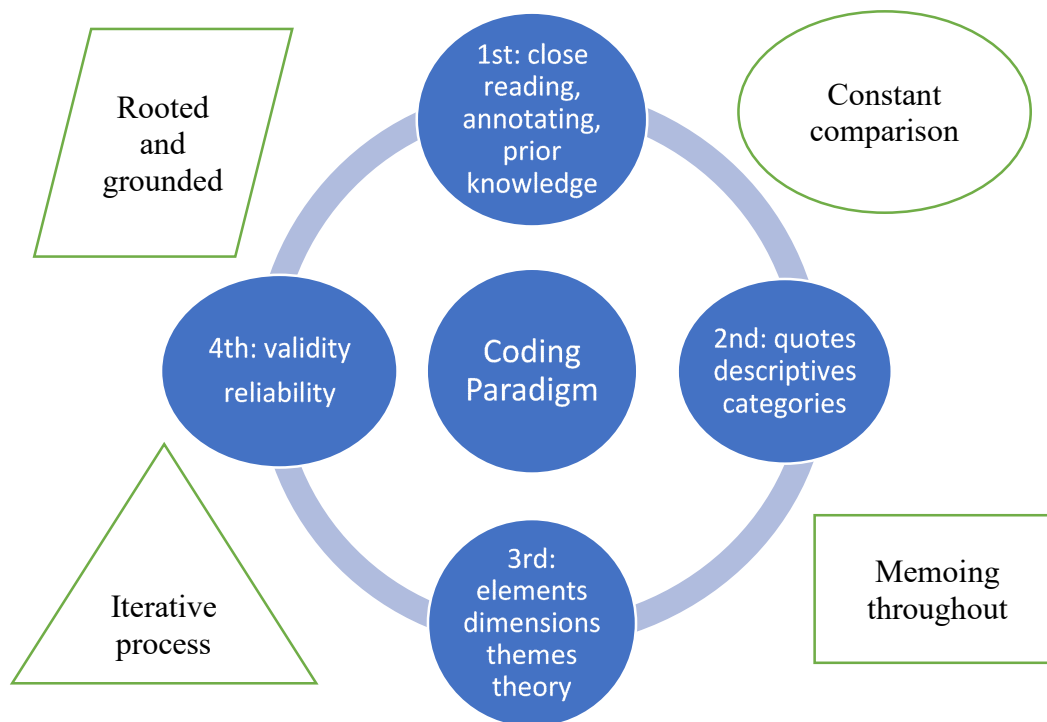


Figure 1. The four stages of the coding paradigm.

Braun and Clarke (2006) say initial coding, but there is little to go on. There are others, such as Nowell et al. (2017) calling for systematic coding, but most textbooks and articles make the entire process mysterious and esoteric. In peer reviews, I heard one followed an article or used a program, and voila, that was all. In my thematic analysis, 14 listed an omnibus method: initial coding, exploratory, and first-order, etc. After 5 stated quoting (or extracts, in vivo, etc.), there was dependency on either grounded theory, descriptive, or close reading (which might mean no coding?). Here was and is the problem: No one stated what initial coding actually looked like. The grounded theory folks had an art, but the others remained mysterious.

A coding schema is a script, such as my four steps, which details what one will do and where one will go. Johnnie Saldaña (2012) provides an excellent resource, but my travails suggest most researchers either do not use systematic coding or they refuse to let everyone in on the secret. When I point out the mystery in my reviews, many authors are thankful and give a detailed report. A paragraph or two in a single study is all that is required. Others refuse, unless the editors will require it. What are we afraid of? I think shoddy work can easily be covered by refusing to explain there is a method in the methodology.

Saldaña (2012) provides much food for thought, but I think we can also be jargon free. One's direction is either deductive/inductive and line-by-line/holistically. First, there should be annotations and readings. In large data sets, there is no other way to keep the information clear. I use the journalism questions of who/what/where/when/why/how, or more generally, context, causes, and consequences (a la Bernard Glaser (1978) in grounded theory). Initially, I do not know what is important, so I overcode. As time passes, and one gets a better feel of the

terrain—for coding is about mapping a data set—one can move faster and become more agile.

Next, quoting (in vivo or extracts) gives a feel for the authenticity. At this point, a geocode needs attached. A geocode could be as simple as numbering each data point (e.g., 1, 2, 3, etc. and attaching the number to each line, or being more specific and having 1.1, 1.2, etc.). A geocode allows all data to be rooted. Next would be a descriptive code, which if one is deductive, might fall into a preordained category. I then like a category code, which might not be possible initially. Memos are thoughts and connections, but I also generate questions or add my attempts at reflexivity. I also have a constant comparison, to formally check and write up thoughts as I move from source to source. There is also an aha! category, to jot down anything which might be important.

After writing up five to seven codes (each with quotes, descriptive, categories, and possibly elements, dimensions, or themes), I write up the themes and my theory about what is going on. I named the process intermittent thematic formation (ITF). I consider the meta-message, or the message behind the message; theming moves to analytic work which considers many factors. Moving through each successive data source allows me to continually revise my thoughts and conclusions. I can also look back and add memos and new thoughts. Reconciliation adds a further layer to the ability to self-audit: Since one never sees the same thing twice, starting back over at the end allows one to make new connections. Rootedness and grounding map the process, and a constant reanalysis can happen.

Some authors recommend a codebook. For small data sets and 60-100 codes, a formal codebook is probably not necessary. For large datasets, a codebook hinges on creating an archival record with indexing. Once three to four data sets have been coded, I promote a key word or phrase in each quote to improve initial indexing. I am using synonyms or similar words/phrases, and I add an asterisk after the word to notate to myself if the key word does not actually appear in the quote. For example, all students and every student were two common phrases, so one was labeled all students and the other one all students* before the original. Some examples are more complex, such as highest potential and full potential. In the end, a judgment was made, and consistency was ensured.

One other problem exists so commonly in peer review, I would be remiss if I did not point it out. Qualitative review should generally have either quotes or the original language in the results section. Hearing and seeing the original words and language give the reader a feel for the data. I am troubled when there are cursory summations without any trace of the actual data. Furthermore, quotes and snippets allow the reader to also make inferences and connections beyond the researcher. I will admit I prefer small snippets of many data sources over a few large quotes; little work is required to drop five or six quotes, but 30 snippets show a sprinkling of the larger picture.

A coding scheme is simply a script to reduce, sort, and expand the data. Coding is labeling and categorizing, and systematically reported and demonstrated codes let the reader look over the researcher's shoulder during the process. Tuckett (2005) and Braun and Clarke (2006), for example, give little guidance here compared to Joffe (2012). Fereday and Muir-Cochrane (2006) provide some details, but they do not say how one actually codes. Other

methodologies, such as grounded theory and phenomenology, do not easily translate into thematic analysis.

#3 *Themata mystici*. Mystical themes.

Apparently, themes just happen. Read some articles. Do initial coding or whatever that means. Sort. Voila. According to my thematic analysis, the mysticism is real. The most common is sorting ($N = 5$) or collating ($N = 4$), though I am not sure what collating even means. Other less common but just as esoteric ways are selecting, mapping, consensus, emerging, generating, eliminating, and developing. *Thou shalt have a method to develop themes.*

If one read the extant literature, one would be lost. The secrecy of researchers continues. From peer review, I am at a loss as well. One either says they followed Braun and Clarke or leave the methods hidden; not one article reviewed gave any detailed explanation. One cannot help but think maybe there was no method—one read and wrote up what sounded good and fit with one's theories and perspectives. It is no wonder researchers like myself default to a concrete methodology, like grounded theory. Because of this lack of transparency, I will tell my story.

There are three interconnected ways to derive themes. First, themes need stages to develop and emerge (more on the emerge part later). Secondly, there is an EVO-DEVO approach, or evolution/development. Finally, a holistic approach must move beyond similarities (Ryan & Bernard, 2003). Themes which are systematic improve validity and reliability, and one must tell how one arrived at the conclusions.

Following the coding schema, themes are integrated into the four steps of coding. There are annotations, codes, and memos initially. All data are data. Next, there is a sorting mechanism, greatly enhanced by categorization; the entire line, from the geocode to the category, is moved next to a category by whatever relationship exists. Once a linkage is made, groupings give a picture of how each one is linked. This initial stage is developing elements, or initial categories. Next, I look for how the elements are linked, and the larger categories are dimensions. Finally, I work to connect the dimensions to a theme. If possible, there might be a larger macrotheme or theory to connect themes. I do not always force all three components of elements, dimensions, and themes, as the richness and depth of the data might preclude such conclusions.

The second way, going on within the sorting, is understanding the EVO-DEVO approach, or evolution/development. There should be a strategy to analyze the data (Humble & Radina, 2018). Coding is about reducing the data in an expansionistic way (splitting). The paradox is not lost, but while the data are reduced, the descriptions and categories grow large initially. Afterward, a development takes the rhizomatic structure back to a foci, or nodule which is large on the element side and reduces again to a dimension and a theme and/or theory.

How does the sorting happen? EVO-DEVO and the formula of elements-dimensions-themes do not tell the complete story. Is there transubstantiation or consubstantiation of the data? How does one distill the data? If one reads the literature, one finds similarities are the most listed method, though some other issues are mentioned. I found we must move beyond only similarities, and as shown in Figure 2, the ROAD is the path: repetition, opposites, absences, and degrees.

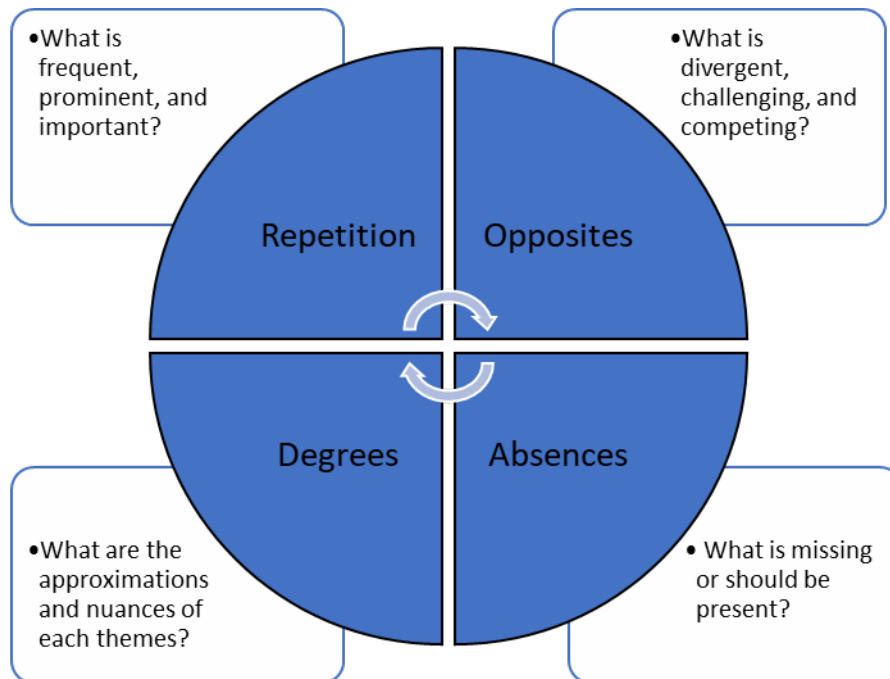


Figure 2. ROAD to develop themes.

There is a distillation to derive linkages between categories and descriptions. The rules of analogies also help: synonyms, antonyms, examples, part to whole, degree, and order, etc. Repetition, or frequency, definitely was front and center in most peer reviews and the thematic analysis, but the researchers who only looks at frequencies and similarities miss a lot. My second thematic analysis, about remote work, used horizontalism, or placing equal importance on all units regardless of frequency. The mission statement analysis considered absences and differences between two samples. Care must be taken to consider and report stubs, or elements, dimensions, and themes which have minor importance but tell an important story.

The degree of each category is a mechanism overlooked within the research on thematic analysis. Themes are never automatic or apparent; if they were, no research would be needed. Decisions must always be made by deciding if one code approximates another, and the coding schema maps the process and allows one to check results are grounded and rooted. A few examples bear this out. In the mission statement paper, full potential and highest potential were coded under the same category. Another example is all students can learn and every student can learn were placed under the same category. Both could have been examined with a different perspective; double coding, especially with the use of frequencies, can inflate one source. “All students can learn in a diverse environment” could be split many ways, though the original intent needs represented in the final analysis.

I hear the ideas for and against emerging. Some claim themes do not emerge. In peer review, authors use emerge, and in my thematic analysis of thematic analysis, it seems to be acceptable. Do themes emerge? Yes. I think an example can make the case: An archaeologist slowly uncovered a corner of an object with a brush. Eventually the corner emerges into a sarcophagus.

Another is uncovering fragments of a container. This example is akin to the researcher using thematic analysis. Is it an urn or a vase? Does one ever definitively know? Another is a scientist mixing two chemicals, which by themselves are disparate but together form a new substance. Still another is an artist writing a book, painting the masterpiece, or the conductor leading the orchestra. All were there before, so did any emerge? One can argue sometimes themes and theories emerge, and other times they are constructed and abstracted. In most cases, the researcher must take the whole, find the parts, and put them back together. This last metaphor should dominate over the pedantic arguments of emerge.

Another problem is claiming thematic analysis is like baking a cake. Baking a cake is a poor metaphor! The cake was already baked. It is up to the researcher to “unbake” the cake and break down the recipe, the development, the pans, the heat, cooling, and the entire process, and then put it back together with insight into how/why the cake was baked. There is difficulty in recreating the process, but the researcher approximates the artifact by developing factors which make up the whole—the EVO-DEVO approach.

Themes must then pass several tests. There should be internal consistency, like the four corners of a contract. There will be overlap, but themes must be well defined to stand on their own. For example, I noticed physical development of students was not in the correct place, so I moved it from physical environment. The entire process is iterative, and one must memo (thoughts, connections, questions, constant comparison, and reflexivity), Aha! (see Saldaña, 2012), and notes. Useful at the end of each source is to write up emergent themes/theories, questions, and start updating with each new source. This rootedness produces a trail which can be audited and show the EVO-DEVO of the entire process as a legitimate representation. Degrees and approximations are necessary components, as coding stages move inductively. Promotion, demotion, and transformations are important concepts, as well as linearity and lateralization. Some themes will be expanded/elaborated, collapsed, or subsumed. Rarely would a theme identified be deleted. Consider the opposite (CtO) gives the researcher a chance to take an outsider’s perspective.

Theming is an active process, and themes can be underdeveloped or overdeveloped (Vaismoradi & Snelgrove, 2019). The results of thematic analysis should be informative and cohesive (Braun & Clarke, n.d.). The researcher is always present in the research. Reflexivity is sometimes mistaken to mean working to keep one’s perspective out of the final results. This reading is mistaken; the researcher’s perspective needs presented, but the perspective needs separated from the results as analysis. A well scripted coding and theming plan provide a systematic way to investigate data and move beyond rapid, first-order decisions to second-order decisions rooted and grounded in the data.

#4 *Sincere proferendi*. Maintain secrecy.

COREQ (Consolidated Criteria for Reporting Qualitative Research) and SRQR (Standards for Reporting Qualitative Research) provide questions as a checklist to improve rigor and trustworthiness of qualitative research (Booth et al., 2014; O’Brien et al., 2014; Tong et al., 2007), yet the continued problems of secrecy and obfuscation continue in much of what is called qualitative research (Anfara et al., 2002). Qualitative research should detail the

methods and decisions involved in the qualitative analysis (Nowell et al., 2017). The confluence of all the other sins is felt here: A visible grounded and rooted report tells the reader the results were not by chance or the whims and desires of the researchers. *Thou shalt have an audit trail.*

The medical professions are ahead of other fields, for they long ago realized standards and formulas improve the accuracy of representations and the resulting analyses. Dependability is the major problem in reliability/validity research. Indeed, the lack of anything more than stating one is systematic is the problem. It is rarely demonstrated beyond stating one followed steps or used a program; the hidden steps remain hidden.

Peer reviewing articles, most authors felt stating the process was superfluous. Unlike in the dissertation, where deviations from approvals by institutional review boards must be stated, many authors omitted any references to the entire research process. There are space limitations, but most authors reacted to my recommendations in two ways. Overwhelmingly, researchers embraced the need for an audit, with a detailed process, quotes, and coding descriptions. The second group, much smaller, refused; one could surmise they were offended, did not feel the time was a worthwhile investment, or they did not have a sustainable method.

Perhaps because of the lack of a visible audit in my peer review and readings, I tried to be very complete within my own work. The more I produced qualitative analysis, the more I realized readers who reviewed work critically needed to know the particulars. The steps advocated by researchers are a starting point; reading thematic analysis, some have pages detailing the process, but most give scant attention. For qualitative research, the problem is rarely mentioned: *methods reproducibility* (Plessner, 2018).

Reproducibility, not replicability, is the major goal of providing a visible trail. Different researchers, from different perspectives and different schools of thought, will and should find different meanings and nuances within the same thematic analysis. What makes good qualitative research from bad qualitative research is another group, using the same sources, could reasonably find the same results. While one could not expect to find the exact same results in another sample, there should be an overlap.

Abductive reasoning drives most qualitative research, so researchers find patterns and groupings which they believe provide the best evidence to support conclusions. The *Canadian Mounted Police Problem* always lurks in every analysis: A researcher always gets their theory. Finding what one wants to find is a real problem; some researchers always find their pet theories and beliefs everywhere they look. The mind is just as good as finding a pattern which exists which does not exist. For said researchers, one knows the conclusion is *always* the same. Adopting reflexivity, by checking the groundedness and rootedness, can help, but researchers need to also consider overtly they are wrong.

Likening qualitative research to a medical doctor assessing a patient, researchers will and do get the diagnosis wrong. An audit trail does not provide a defense to such mistakes, but providing a detailed method gives the reader the chance to consider alternatives which have plausibility and can be connected to the evidence. Without the trail and evidence, though,

readers are stuck with little to reconsider, reimagine, and theorize.

#5 *Argumentum ad verecundiam*. The argument is right because the researcher says so.

Most articles reviewed make wild claims, such as “ensured trustworthiness” or “reached saturation.” Validity has many angles and issues, making simple pronouncements irrelevant (Norris, 1997). Such claims are probably untrue and counter to a postpositivistic view, which by its very nature means results are not positivist. How or what was either ensured or saturated is a nameless process. *Thou shalt have reliability and validity*.

Reliability and validity are connected to others’ reformulations around trustworthiness (and the different criteria), triangulation, saturation, and legitimation/representation. All the components have a significant overlap, but the more I examine and produce my own research, the more I believe the different criteria are needlessly complex. Forero et al. (2018) probably do a better job than any other at making the components of reliability and validity operational, as outlined by Lincoln and Guba (1986). The whole issue of reliability and validity can be recast as one question: Do the results accurately and plausibly represent the data? Four elements can improve reliability and validity: systematization, triangulation, rootedness, and groundedness.

Systematization connects all processes of reliability and validity. Many components need considered: sample selection and size, design, audit trail, and check for rootedness and groundedness. Reconciliation is a way to check for reliability and validity. I found recoding after a time lapse in larger studies meant if different perspectives or codes emerged, the new codes could be reconciled with the old codes. How or why could the same material be recoded differently? It could be I was wrong, but I found as I dug deep into a topic, my perspectives and thoughts grow with the coding, reviewing research, and constant comparison. I overtly considered the opposite of what I thought to try to keep my own Canadian Mounted Police Problem at bay.

Onwuegbuzie et al. (2008) point out bias does not need rooted out. Bias, as stated by Kahneman et al. (2021), might not be the only problem: noise. Noise is not bias but effects the decision-making process. Researchers need to write what is socially acceptable and academically expected, else the work might not be published. The inherent randomness in much of what qualitative research will be is cause for concern from both bias and noise, so providing a well-defined and well-explained methodology can reduce problems which challenge reliability and validity.

Many researchers claim thick descriptions increase trustworthiness. I sheepishly admit I have jumped on this bandwagon. I believe claiming thick descriptions increase trustworthiness is often spurious; anyone can be imaginative and thoughtful about findings, but how thick descriptions increase plausibility and representation remain doubtful unless rooted and grounded. Doubtlessly, many researchers have been ensnared by the thick description trap: Write lots of bull and dissect to many angles, and one uses Campbell’s Law to game the trustworthiness requirements.

Triangulation, by considering data, theories, researchers, and methodologies, can be

implemented. Some authors, in peer review, my thematic analysis, and my own work, use training sets. Training sets are data sources which have been excluded to test for new major ideas. There is the fear researchers will be weary and find what they always have, but an explicit examination for dysfluency and divergence can help control for biases. Some divergence should be expected, but a future problem could be the lack of reporting one's initial themes did not work out (since the initial themes always work, there is the probability of the improbable, that everyone *always* gets everything right the first time).

Saturation. Saturation is often the gold standard of reliability and validity. Fusch and Ness (2015) offer the advice if new information continues to be mined, then more interviews should be present. Still, saturation remains esoteric and difficult to define, with multiple, often competing definitions (Lowe et al., 2018). Rarely does saturation direct sample size selection (W. C. Morse et al., 2014), though there are recommendations for minimum sizes in interviews (e.g., Ando et al. (2014), Francis et al. (2010), etc.).

The thematic analysis found few researchers mentioned saturation (4 of 35, with all dealing with interviews, and claims were rarely justified; the claims of theoretical saturation when developing themes sound dubious). Within my own research, I think saturation is real and felt. After three or four articles, a noticeable pattern developed within the data. The sense was like *déjà vu*, with the feeling I was reading or hearing the same thing. Yet, like authors in the thematic analysis, saturation is inappropriate if one is using a frequentists approach with codebooks or a well-defined sample, such as scoping reviews.

Hypothetically, if a different researcher followed the same defined path of a published article, the person would find and see differences. That does not negate the other researcher as long as the researcher could also reasonably find the results published. I throw around reasonableness and plausibility, but that factor should be celebrated with the postpostivist paradigm instead of decried. Researchers who do not reach saturation are not necessarily any less reliable and valid than authors who do (O'Reilly & Parker, 2013).

When does saturation happen? To alter a quote from the American jurist Oliver Wendell Holmes, one will know saturation when one sees it. Other authors freely admit even after saturation, new ideas can emerge and be developed, but major themes will not be present. A good recommendation, like in the thematic analysis, my peer review, and my research, is saturation occurs in gradations, and reporting it seems to offer little *unless* one can claim there are no major new developments. Fabrication and false coherences are real problems researchers face (Fereday & Muir-Cochrane, 2006), but the best researchers can hope is to show the methodologies and methods to let readers develop skepticism. How does one know this unless there are large samples and multiple (truly independent) coders?

#6 *Perfectio ad arrogantia*. The pride of perfection.

Limitations abound in every study. Using the What Works Clearinghouse, one can center limitations around design, attrition (if applicable), baseline, confounders, and outcomes. Two of the factors, attrition and baseline, deal with sampling and are generally not applicable to content analysis unless sources have been overlooked. In my peer reviews, I frequently find

authors state no limitations. In the thematic analysis, 6 of 35 achieved perfection—there were no limitations. *Thou shalt have limitations.*

Within the thematic analysis, sample dominated the limitations (36 times samples were mentioned in limitations across 17 articles). There were four areas: skewed, small size, method, and selection. Vasileiou et al. (2018) points out the problem in trying to apply quantitative rules to qualitative research. Small samples can be perfectly acceptable if justified, but care must be taken to maintain anonymity and confidentiality (J. Morse, 2008). That does not negate the limitation in most samples; there might be redundancy, waste of time with so little new information after a certain point, and lack of resources, but one always wonders (or should) what would happen if interviews went from 12 to 120.

The other reasons stated in the thematic analysis were the following: data analysis, data collection, subjectivity, lacking member check, lacking reflexivity, one researcher, and did not reach saturation. Limitations can be examined from the hypothetical: If I had to do the study over, what would I have done differently? What was missing? There is always something missing. Limitations speak to the possibilities not followed, and when explained, give readers an improved sense of not only generalizability but also reproducibility.

#7 *Cohaesionis*. Cohesion.

The last problem is pervasive: extreme cohesion. Conducting peer review, I see 100% coherence way too much. I have a rule to evaluate thematic analysis: If there are no divergences, outliers, and/or more questions/unknowns, then the research lacks credibility (Patton, 1999). Reconciliation and reevaluation always bring up new nuances and redefine what I think I knew. *Thou shalt have divergence.*

Research is messy, and unless there is extreme homogeneity—to the point one might as well have a sample of one—there should be disagreements, dysfluencies, and doubts. The stubs or miscellaneous are common occurrences within any data set. I have seen samples of 400 with no reports of variances. Variances are important, and if one is not reporting them, there might be more than shoddy work: theme blindness. Once someone gets a pattern in their mind, and unless one actively pushes out the dominant narrative and considers different perspectives, it is easy to find what one wants to see.

My method is to consider the opposite. Bad faith, as mentioned by Sarte, might be real and at play within any context, as politics can drive data. Though derided in the popular media, lawyers look for alternative facts, or as Yin (2017) would say, disconfirmation. The word alternative should not scare one away; there are facts which can be organized differently than the dominant story. Life and research are messy, and lack of cohesion should be the cohesion which holds the analysis together.

Limitations

There were some limitations to the study. The sample, both from the autoethnography and the thematic analysis, is not especially large and does not map the entire thematic analysis literature. Autoethnography can be very useful, but it relies on the perspective of memory and

review of artifacts (Ellis et al., 2011; Wall, 2006). I struggle, like most people, of casting myself as the hero in a narrative. All narratives reduce and focus on some things and not others. Finally, the thematic analysis was primarily descriptive in nature; a different approach might yield more useful analysis. How to develop themes with relationships which go beyond similarities, such as moderation and mediation, remain to be utilized.

Conclusion

Thematic analysis, and qualitative research in general, is akin to prosopography, where broader social connections are the goal through the aggregation of the disparate data sources. External validity—and usefulness—depends on our ability to distill common, defining characteristics and connect them to the greater temporal and spatial epoch of the world. All qualitative research seeks to make visible the endogenous and exogenous changes in society, and making the esoteric real and knowable make thematic analysis useful and applicable.

Thematic analysis is a messy, iterative process with continuously evolving methods throughout. One should have a well-developed methodology upfront, and one must document the changes which will happen as one encounters data and analysis. Narratives try to create verisimilitude and minimize falsifications; there will be themes and components which are false, but some are false than others. Formally turning on a critical posture and searching for confounders can improve the process while also honoring the inherent lack of coherence. The rootedness and groundedness can help to evaluate the continuum of the truthiness; there should be a manifest weight of the evidence. This is why reproducibility, not replicability matters. Different researchers might see things differently, but a new set of participants will be at a different point in space and time.

Kahneman et al. (2021) pointed out simple rules and models outperform individual judgments. Researchers deal with the facticity of a matter, and one can easily suffer from the *unicorn problem*. The unicorn problem is driven by the need to be complicated, innovative, and novel in order to be published. The unicorn is usually adorned in one's pet theory. A hashtag should follow researchers: #formulaicflexibility. Simplicity, rooted and grounded in the data, should be craved by researchers with the strong outline of the use of systematization.

There is the claim postpositivism suffers from a crisis, but the crisis exists on the positivist side just as much. While positivist do not have to explain statistical methods due to standardization, there are other shortcomings. I reviewed a paper which *never* explained the variables, stating see five different articles. How anyone could review the entire paper without reading 80 pages is beyond me. Another issue is many authors only report p-values; giving complete statistics, assumptions, and a detailed sample report allow the reader to interpret and generalize results. Two other major problems exist on both sides. Rarely do I see divergence, or everything is always statistically significant. Impossible. Secondly, any result, no matter how small, is the crux of the article. A p-value of .051 does not matter, but .05 does. Another issue is authors either do not state a baseline (making any comparison impossible) or make any finding, no matter how small, important.

There will be sins, errors, and omissions in any research. Following virtues are a guide, and when we fall short, we must report our shortcomings for the world to see. Researchers on both sides of the aisle of positivism and postpositivism are not used to showing such weaknesses, but astute readers profit from being able to look over the shoulders of researchers.

References

- Anderson, L. (2006). Analytic autoethnography. *Journal of Contemporary Ethnography*, 35(4), 373-395. <https://doi.org/10.1177%2F0891241605280449>
- Ando, H., Cousins, R., & Young, C. (2014). Achieving saturation in thematic analysis: Development and refinement of a codebook. *Comprehensive Psychology*, 3(03), CP.3.4. <https://doi.org/10.2466/03.CP.3.4>
- Anfara Jr, V. A., Brown, K. M., & Mangione, T. L. (2002). Qualitative analysis on stage: Making the research process more public. *Educational Researcher*, 31(7), 28-38. Retrieved from <https://files.eric.ed.gov/fulltext/ED454243.pdf>
- Attride-Stirling, J. (2001). Thematic networks: An analytic tool for qualitative research. *Qualitative Research*, 1(3), 385-405. <https://doi.org/10.1177%2F146879410100100307>
- Booth, A., Hannes, K., Harden, A., Noyes, J., Harris, J., & Tong, A. (2014). COREQ (consolidated criteria for reporting qualitative studies). *Guidelines for reporting health research: A user's manual*, 214-226. John Wiley & Sons.
- Braun, V., & Clarke, V. (n.d.) *Thematic analysis: A reflexive approach*. University of Auckland. <https://www.psych.auckland.ac.nz/en/about/thematic-analysis.html>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101. <https://doi.org/10.1191/1478088706qp063oa>
- Creswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative*. Prentice Hall.
- Ellis, C., Adams, T. E., & Bochner, A. P. (2011). Autoethnography: an overview. *Historical Social Research/Historische Sozialforschung*, 273-290. Retrieved from <http://nbn-resolving.de/urn:nbn:de:0114-fqs1101108>
- Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, 5(1), 80-92. <https://doi.org/10.1177%2F160940690600500107>
- Forero, R., Nahidi, S., De Costa, J., Mohsin, M., Fitzgerald, G., Gibson, N., ... & Aboagye-Sarfo, P. (2018). Application of four-dimension criteria to assess rigour of qualitative research in emergency medicine. *BMC Health Services Research*, 18(1), 120. <https://dx.doi.org/10.1186%2Fs12913-018-2915-2>
- Francis, J. J., Johnston, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M. P., &

Grimshaw, J. M. (2010). What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology and Health, 25*(10), 1229-1245. <http://dx.doi.org/10.1080/08870440903194015>

Fusch, P. I., & Ness, L. R. (2015). Are we there yet? Data saturation in qualitative research. *The Qualitative Report, 20*(9), 1408. Retrieved from <http://www.nova.edu/ssss/QR/QR20/9/fusch1.pdf>

Glaser, B. (1978). Theoretical sensitivity. *Advances in the methodology of grounded theory*. The Sociology Press.

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods, 18*(1), 59-82. <https://doi.org/10.1177%2F1525822X05279903>

Guest, G., Namey, E., & Chen, M. (2020). A simple method to assess and report thematic saturation in qualitative research. *PLoS One, 15*(5), e0232076. <https://doi.org/10.1371/journal.pone.0232076>

Humble, Á., & Radina, E. (Eds.). (2018). *How qualitative data analysis happens: Moving beyond "themes emerged."* Routledge.

Joffe, H. (2012). Thematic analysis. In D. Harper and A. R. Thompson (Eds.), *Qualitative research methods in mental health and psychotherapy* (pp. 209-223). John Wiley & Sons.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little Brown Spark.

Lincoln, Y. S., & Guba, E. G. (1986). But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. *New Directions for Program Evaluation, 1986*(30), 73-84. <https://doi.org/10.1002/ev.1427>

Lowe, A., Norris, A. C., Farris, A. J., & Babbage, D. R. (2018). Quantifying thematic saturation in qualitative data analysis. *Field Methods, 30*(3), 191-207. <https://doi.org/10.1177%2F1525822X17749386>

Lucas, K., & D'Enbeau, S. (2013). Moving beyond themes: Reimagining the qualitative analysis curriculum. *Qualitative Communication Research, 2*(2), 213-227. Retrieved from <https://ir.library.louisville.edu/cgi/viewcontent.cgi?article=1350&context=faculty>

Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample size in qualitative interview studies: Guided by information power. *Qualitative Health Research, 26*(13), 1753-1760. <https://doi.org/10.1177/1049732315617444>

Morse, J. M. (2008). "What's your favorite color?" Reporting irrelevant demographics in qualitative research. *Qualitative Health Research, 18*(3), 299-300. <https://doi.org/10.1177/1049732307310995>

Morse, W. C., Lowery, D. R., & Steury, T. (2014). Exploring saturation of themes and spatial locations in qualitative public participation geographic information systems research. *Society*

- & *Natural Resources*, 27(5), 557-571. <https://doi.org/10.1080/08941920.2014.888791>
- Norris, N. (1997). Error, bias and validity in qualitative research. *Educational Action Research*, 5(1), 172-176. <https://doi.org/10.1080/09650799700200020>
- Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, 16(1), 1609406917733847. <https://doi.org/10.1177%2F1609406917733847>
- O'Brien, B. C., Harris, I. B., Beckman, T. J., Reed, D. A., & Cook, D. A. (2014). Standards for reporting qualitative research: a synthesis of recommendations. *Academic Medicine*, 89(9), 1245-1251. <https://doi.org/10.1097/ACM.0000000000000388>
- Onwuegbuzie, A. J., Leech, N. L., & Collins, K. M. (2008). Interviewing the interpretive researcher: A method for addressing the crises of representation, legitimation, and praxis. *International Journal of Qualitative Methods*, 7(4), 1-17. <https://doi.org/10.1177%2F160940690800700401>
- O'Reilly, M., & Parker, N. (2013). 'Unsatisfactory saturation': a critical exploration of the notion of saturated sample sizes in qualitative research. *Qualitative Research*, 13(2), 190-197.
- Patton, M. Q. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Services Research*, 34(5 Pt 2), 1189-1208. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1089059/>
- Plesser, H. E. (2018). Reproducibility vs. replicability: a brief history of a confused terminology. *Frontiers in Neuroinformatics*, 11, 76. <https://doi.org/10.3389/fninf.2017.00076>
- Ryan, G. W., & Bernard, H. R. (2003). Techniques to identify themes. *Field methods*, 15(1), 85-109. <https://doi.org/10.1177/1525822X02239569>
- Saldaña, J. (2012). *The coding manual for qualitative researchers*. Sage.
- Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, 8(1), 1-10. <https://doi.org/10.1186/1471-2288-8-45>
- Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*, 19(6), 349-357. <https://doi.org/10.1093/intqhc/mzm042>
- Tuckett, A. G. (2005). Applying thematic analysis theory to practice: A researcher's experience. *Contemporary Nurse*, 19(1-2), 75-87. <https://doi.org/10.5172/conu.19.1-2.75>
- Vasileiou, K., Barnett, J., Thorpe, S., & Young, T. (2018). Characterising and justifying sample size sufficiency in interview-based studies: Systematic analysis of qualitative health research over a 15-year period. *BMC Medical Research Methodology*, 18(1), 1-18. <https://doi.org/10.1186/s12874-018-0594-7>
- Vaismoradi, M., & Snelgrove, S. (2019). Theme in qualitative content analysis and thematic

analysis. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 20(3).
<https://doi.org/10.17169/fqs-20.3.3376>

Wall, S. (2006). An autoethnography on learning about autoethnography. *International Journal of Qualitative Methods*, 5(2), 146-160.
<https://doi.org/10.1177%2F160940690600500205>

Yazan, B. (2015). Three approaches to case study methods in education: Yin, Merriam, and Stake. *The Qualitative Report*, 20(2), 134-152. Retrieved from <http://nsuworks.nova.edu/tqr/>

Yin, R. K. (2017). *Case study research and applications: Design and methods*. Sage Publications.