Paquibot, Jesusa L. et al.

Working Paper

Using the framework method for the analysis of triadic interview data: Process notes from a curriculum review

PIDS Discussion Paper Series, No. 2025-24

Provided in Cooperation with:
Philippine Institute for Development Studies (PIDS), Philippines

Suggested Citation: Paquibot, Jesusa L. et al. (2025) : Using the framework method for the analysis of triadic interview data: Process notes from a curriculum review, PIDS Discussion Paper Series, No. 2025-24, Philippine Institute for Development Studies (PIDS), Quezon City

This Version is available at:
https://hdl.handle.net/10419/331367

DISCUSSION PAPER SERIES
No. 2025-24, September 2025

Photo from the Philippine News Agency (https://www.pna.gov.ph/photos/72533)

EDCOM2
The Second Congressional Commission on Education

Using the Framework Method for the Analysis of Triadic Interview Data: Process Notes from a Curriculum Review

Jesusa L. Paquibot, Charlotte Marjorie L. Relos, Edmar E. Lingatong, Erwin Doroteo Justien C. Daga, Tania Dew S. Perez, and Michael R.M. Abrigo

This study is carried out with support from the Second Congressional Commission on Education (EDCOM II).

CONTACT US:

RESEARCH INFORMATION DEPARTMENT
Philippine Institute for Development Studies

18th Floor, Three Cyberpod Centris - North Tower
EDSA corner Quezon Avenue, Quezon City, Philippines

https://www.pids.gov.ph
publications@pids.gov.ph
(+632) 8877-4000

Using the Framework Method for the Analysis of Triadic Interview Data: Process Notes from a Curriculum Review

Jesusa L. Paquibot, Charlotte Marjorie L. Relos, Edmar E. Lingatong, Erwin Doroteo Justien C. Daga, Tania Dew S. Perez, and Michael R.M. Abrigo

PHILIPPINE INSTITUTE FOR DEVELOPMENT STUDIES

JULY 2025

Abstract

We document and reflect on modifying the framework method for the qualitative data analysis (QDA) of large-scale triadic interviews with education actors. The QDA is part of a curriculum revision's early assessment and impact evaluation, which used a mixed methods design including a teacher survey, key informant interviews, and focus group discussions. The participants work in schools sampled for the assessment using cluster randomized design that assigned schools to either the treatment or control curriculum. Considering the amount of interview data and the diversity of our team, we aimed for a very high intercoder and intracoder reliability (ICR) in creating and applying our analytical framework or coding tree during data condensation and analysis. Clearly defining the codes with inclusion and exclusion boundaries, segmenting the transcripts, and memoing were the key processes that increased our ICR. Assessing the codes application and performance using Fleiss Kappa improved the working analytical framework's communicability, consistency, and transparency. Memoing helped trace the points of interpretive divergence among the coding team. After coding, we indexed the segmented data in a framework matrix, using relatively more accessible or free tools to organize, display, and interpret the data. With these procedures, we aimed to meet the consolidated standards for reporting qualitative data (COREQ).

Keywords: framework method, mixed methods research, intercoder reliability, intracoder reliability, curriculum revision

Table of Contents

List of Tables

List of Figures

Using the Framework Method for Analysis of Triadic Interview Data: Process Notes from a Curriculum Review Application

Jesusa L. Paquibot[1,☼], Charlotte Marjorie L. Relos[1], Edmar E. Lingatong[1], Erwin Doroteo Justien C. Daga[2], Tania Dew S. Perez[3], and Michael R.M. Abrigo[4]

1. Introduction

Developed in the 1980s for large-scale social policy research, the framework method (FM) is a step-by-step guide for doing content and thematic analysis of interview data (Gale et al. 2013; Collaco et al. 2021; O'Connor and Joffe 2020), including those from participants selected using a randomized controlled trial as this paper demonstrates. As FM systematizes an audit-trail of the qualitative data analysis, teams composed of multiple and fairly heterogeneous researchers may find it useful in qualitative data analysis, particularly building evidence that could be cross-referenced when challenging or substantiating quantitative results. In this paper, we document and reflect on how we used FM's stages for an assessment using a mixed methods approach. The study context is the first phase of implementing a curriculum revision in Philippine public schools in the school year 2023 to 2024. The results of the study are discussed in a separate monograph where we triangulate quantitative and qualitative evidence to make an early assessment of the curriculum revision (Abrigo et al. forthcoming).

While the literature documents the FM stages, the applications we referenced as guides were mostly in health and social care context (Gale et al. 2013; Klingberg et al. 2024; Arifin et al. 2019). For example, Collaco et al.'s (2021) application interviewed young men affected by cancer and their partners to understand their experiences and needs. Collaco, et al. shared how FM can be used for dyadic interview analysis to surface the multiple perspectives of couples as well as the individuals within the couple (Gale et al. 2013; Klingberg et al. 2024). Continuing this approach, our application builds analytical consensus for triadic data and adds to the literature of applying FM in education research.

Teams practicing collective analysis where multiple researchers analyze data need to reach a consensus on data interpretation. In the framework method, consensus is checked by assessing intercoder and intracoder reliability (O'Connor and Joffe, 2020). Intercoder reliability measures how consistently the codes are used when indexing and coding segments of the interview transcripts. Intracoder reliability measures how similarly the data is coded in different periods. In this study, ICR was assessed using Fleiss Kappa. Evaluating ICR at strategic stages of the framework method, such as after open coding and the creation of the initial version of the working analytical framework, ensures that coders have a shared understanding and application of the codes, leading to a common interpretation of the data. The ICR provides a documented record of the data analysis process, facilitating transparency and researcher reflexivity (O'Connor and Joffe 2020).

[1] Project technical specialist, Philippine Institute for Development Studies (PIDS)
[2] Project technical assistant, PIDS
[3] Research analyst, PIDS
[4] Senior research fellow, PIDS
☼Corresponding author

Aiming for high intercoder reliability, we decided to segment and review the segmenting of transcripts before applying our analytical framework. Segmenting transcripts fixes the coders' analytical frame, helping increase consensus on how to use the codes. Like other qualitative data analysis methods, researchers using FM must consider meeting the COREQ standards or practices that increase reliability and transparency of the data condensation and interpretation, as well as ethical guidelines. We find that FM can meet the COREQ standards, and in this paper discuss how we conducted researcher reflexivity practices throughout the FM stages using memoing. In the next sections, we discuss the framework method and the context of our application. We discuss how we conducted the method and the reasons for our modifications as we encountered application challenges in assessing a curriculum revision. We then describe the activities in the method's stages and share practical notes for researchers who wish to unpack the underlying method used in computer-assisted QDA tools (Ritchie et al. 2003) or those who want to apply the method "manually" using relatively more accessible tools.

2. The framework method

The framework method consists of stages or steps done to "systematically develop and reduce data to analyze it by case and code" (Collaco et al. 2021). The steps are shown in Table 1, with Gale et al.'s (2013) version as the basic mode of conducting the method. Arifin et al. (2019) noted that the framework method can be modified according to the study purpose and standards that the researchers want to meet, yet its overall process can still be followed. Table 1 describes how three studies have been clustered to meet the researchers' objectives. Some stages can be added or moved to earlier or later parts of the sequence. For our modification, we referred to the consolidated criteria for reporting qualitative research (COREQ) in making decisions for modifying the framework method (Tong et al. 2007) and made methodological memos to document and reflect on our use of the method. Our modification aimed for higher ICRs, which was gauged using a sample of transcripts. Only when we had high ICRs did we apply the codes to the rest of the transcripts.

The method facilitates the thematic analysis of hundreds of interviews by organizing the stories in the transcripts using an analytical framework. The stories found in the transcripts are summarized as a coding tree consisting of themes, major codes, and minor codes. The analytical framework is iteratively developed throughout the FM stages as a tool for data reduction and organization. In our application, each code has a definition that connotes multiple directions (positive, negative, neutral responses), as well as definitional boundaries or exclusion and inclusion criteria.

The FM's output is a coding matrix, which contains rows for cases and columns for codes and representative statements. The matrix organizes the views of the participants in one table, facilitating the constant comparative approach across and within groups. Collaco et al. (2021) noted that displaying the data this way allows for transparency showing how the findings are derived. The text-and-context dynamics are also considered through this method as "the participants' views remain connected to other aspects of their account within the themed matrix" (Collaco et al. 2021, p. 1557).

Collaco et al. (2021) used the framework method to investigate the multiple perspectives gained from analyzing separate interviews from the point of view of the dyad (or the pair taken as a whole) with interest on the overlaps and variations in each participants' version. As mentioned, the stories in this study were told in the context of the FGDs where teachers could qualify, disregard, agree, or disagree with each other and the KIIs where insights from school or division leaders could be discussed. Analyzing triadic interview data from the FGDs and

KIIs showed convergence and divergence in the participants' views. One participant group might claim one thing while another group makes a slightly different or opposite claim, prompting further investigation on the relevant concern and providing a more nuanced understanding of a theme by describing the range of the participants' experiences.

Table 1. FM stages clustering in previous studies vs this FM's clustering

Gale et al. (2013)	Arifin et al. (2019)	Collaco et al. (2021)	This process
Stage 1: Transcription	Stage 1: Data management and transcription selection	Stage 1: Transcription	Stage 1: Transcription
Stage 2: Familiarization with the interview		Stage 2: Familiarization	Stage 2: Familiarization
Stage 3: Coding		Stage 3: Coding	Stage 3: Coding Open coding Question-based coding
Stage 4: Developing a working analytical framework	Stage 2: Descriptive accounts	Stage 4: Charting codes into table of themes	Stage 4: Developing a working analytical framework Feature-based coding
		Stage 5: Dyadic analysis	Stage 5: Segmenting and reviewing segments
		Stage 6: Developing a working analytical framework	
Stage 5: Applying the analytical framework		Stage 7: Applying analytical framework	Stage 6: Tagging or applying the analytical framework
Stage 6: Charting data into the framework matrix (entering summarized data into the matrix)	Stage 3: Explanatory accounts		Stage 7: Charting data
Stage 7: Interpreting the data		Stage 8: Interpreting data	Stage 8: Interpreting the data; triadic analysis

3. Application context

In 2023, the Department of Education (DepEd) launched the pilot of a revised curriculum dubbed as MATATAG. According to DepEd (2024, par. 6),

> MATATAG stands for (1) Make the curriculum relevant to produce competent, job-ready, active, and responsible citizens; (2) Take steps to accelerate the delivery of basic education facilities and services; (3) Take good care of learners by promoting learner well-being, inclusive education, and a positive learning environment; and (4) Give support to teachers to teach better.

We used a mixed methods design to assess the first year of implementing the curriculum in Kindergarten and Grades 1, 4, and 7. We combined qualitative data from interviews with classroom observation, a curriculum design review by subject experts and quantitative data from a teacher survey and a student assessment. The interview participants work in schools

assigned to control and treatment groups in the school year 2023 to 2024. Sampling of schools was done using a cluster randomized design where blocks were assigned per region. The randomized controlled trial considered school level, size, and urbanicity. Indigenous people's schools with similar characteristics were also randomly sampled.

The team conducted seventy FGDs with teachers (N=686) assigned to kindergarten and grade levels 1, 4, and 7 or the first phase of the MATATAG implementation. Ninety-seven principals or school heads and 91 administrators participated in key informant interviews. The semi-structured interviews had seven main topics, namely curriculum objectives, classroom instruction, learning resources, professional development, learning assessment, mother tongue-based multilingual education, teacher workload, and satisfaction. Participants talked about prospectives and retrospectives on the two curricula. Some topics were discussed with all the participant groups; others were tackled if appropriate to their role. We analyzed over 160 transcripts, each with about 1.5 hours of interview time. We used transcripts in the local languages we could understand, although some transcripts translated into English were used depending on the coder's preference.

Our team members varied in terms of age, educational levels, and disciplinary fields in the Social Sciences (see Table 1). We had some training and experience in either quantitative or qualitative research but most of us were first-time users of mixed methods research and the framework method. Three team members took part in conducting the interviews. One member was a coordinator involved in transcribing some of the KIIs and FGDs prior to the data analysis. One team member had not encountered the participants in person before the data analysis.

Table 2. Research team composition

Team	Field	Educational level	Age	Sex
Researcher 1	Economics	PhD	38	M
Researcher 2	Sociology	BA, with MA units	31	F
Researcher 3	Economics	MA	26	F
Researcher 4	Geography	BA, with MS units	25	M
Researcher 5	Economics	BA	24	M

Team composition varied per FM stage, but two researchers were mainstays throughout the whole process. At least three researchers created and revised the analytical framework because having more perspectives can foster better and faster consensus-building (Gale et al, 2013). Three to four researchers were involved in the ICR tests. Once we reached a satisfactory ICR, the team composition was reduced during segmenting and tagging. Throughout these stages, one advanced researcher provided feedback on the process and outputs. Their perspective was crucial in setting standards for the usability and reliability of the coding framework. The framework matrix (extraction using MS Word and charting data using Stata and Excel) was done by two members. Three researchers distributed the analysis of each code for the last stage (writing the findings). As an interdisciplinary research team with varying experience levels in qualitative analysis, we found the FM useful in meeting the COREQ standards as it provides clear means of assessing the quality of our team's analysis. The method helped us to stay close to the data as its systematic approach provides a more manual building of qualitative evidence from large-scale interviews.

4. Procedure of analysis and resources required

We first inductively constructed the coding tree through line-by-line or open coding of the transcripts. Then, the initial codes were grouped into categorical codes which were, in turn,

grouped into analytical codes. To develop a comprehensive yet succinct coding tree, we switched to a more deductive approach, preparing codes guided by our questions. Finally, we relied on the related literature while using a combination of inductive and deductive coding to finalize our coding tree. Also known as an analytical framework, a coding tree classifies all codes into a framework. Each code summarizes stories, topics, and/or themes from the interviews. Clear boundaries for each code are specified through inclusion and exclusion criteria. Our process followed the usual FM stages but used pre-segmented interview transcripts in the coding stage (applying the analytical framework). After the codes were applied to all the segmented codes, we used a framework matrix to analyze the data to trace how evidence was used to write the findings.

The framework method contains the four steps of thematic analysis: data immersion, coding, creating categories, and identifying themes (Green et al. 2007). Through FM, more researchers can participate in thematic analysis as they can easily compare the organized data using the framework matrix. On the other hand, content analysis practitioners may synthesize interviews by describing its content in terms of frequency, occurrence, and patterns using the coding framework (O'Connor and Joffe 2020). In this way, the QDA can help validate survey data by supporting or qualifying relationships and trends.

For this section, we provide activities per stage, resources used, and lessons learned from implementing the method. It is followed by a reflection on the effect of segmenting on the quality of analysis. The framework analysis does not necessarily make qualitative research easier. Researchers should plan enough time to go through the stages and create their own modifications to meet their study requirements. In Table 3, we provide estimates of the time needed to conduct the method. We worked on the coding tree and achieved a high ICR for about 3 months using sample transcripts, and segmented and tagged the rest of the transcripts for another 3 months.

Stage 1: Transcription

Four to five coordinators transcribed the interview data for about a month for each region (12 to 14 schools). They transcribed the audio recordings verbatim, following a format that noted timestamps and anonymized the participants. The transcriptions were prepared following the formatting suggested in the literature: Times New Roman, 12, wide margins, and double-spaced for the coders' notes, and continuous line numbers. While this format was helpful in the initial parts of the study, the most useful were the line numbers which we often referred to during the workshops. We used transcript codes indicating the School/Division code, Block or PSGC code, Respondent group, and Treatment group. For example, the transcript document "S21PP", pertains to school number assigned based on block and PSGC, respondent group, and treatment group. Interview data were stored in a password-protected drive that only the research team could access.

The transcripts were distributed randomly among the researchers or coders and thus, a researcher mainly encountered the prepared transcripts and thus coding relied heavily on what was said and documented during the interview. Taking care of context-and-text dynamics when reading transcripts was prioritized during analysis. Helping contextualize the reading of the conversation, the transcribers indication of laughter, such as to note jokes or sarcasm were considered in our reading and highlighting of transcripts. Hesse-Biber (2017) suggested conducting memos while preparing and familiarizing with the transcripts as it already begins the thematic analysis.

Stage 2: Familiarization
Familiarization involves getting a general sense or impression of the data, highlighting the analytically relevant parts of the transcripts and making memos or preliminary labels or codes. Two researchers immersed in the first batch of available transcripts from one region. At this stage, we read the transcripts as is and did not yet refer to the relevant literature on the curriculum revision when labelling and highlighting segments. Individual researchers prepared notes or memos of initial impressions, insights and suggested codes to discuss with the team in the next stage, where transcripts from the rest of the regions were sampled to verify the initial codes emerging from the first batch of transcripts.

Stage 3: Coding
The coding process had three phases of regular workshops aimed at enhancing the usability of the coding tree. It began with inductive or open coding, progressed to deductive or question-based coding, and ended in a phase where we aimed to strike a balance between inductive and deductive coding.

Coding Phase One: Open coding (Inductive)
According to Gale et al. (2013), "codes could refer to substantive things, values, emotions, and more impressionistic or methodological elements" (p. 4). In open coding, we highlighted and coded anything that "might be relevant from as many different perspectives as possible" (Gale et al. 2013). Gale et al., noted that it is "generally helpful even if you are taking a broadly deductive approach to do some open coding on at least a few of the transcripts to ensure important aspects of the data are not missed." Open coding helps surface the unexpected by alerting researchers to the invisible, which is crucial when encountering related literature before coding. Gale et al. also noted the value of involving a heterogeneous researcher team so they can "offer alternative viewpoints... ensuring that one particular perspective does not dominate," and to ensure that the coding can go beyond the literal and descriptive, toward more analytical coding.

At least two researchers can start the process using a sample of the transcripts. Allocating about two weeks of focused work on open coding can yield a huge number of initial codes, which are the starting points for constructing the analytical framework by grouping the initial codes into categorical codes, analytical codes, and thematic codes (Hesse-Biber 2017). In our practice, we had 400 descriptive codes for the seven to eight main topics or blocks of the interview guide. During discussions, we agreed on the final naming of the codes and on how to group them into categorical codes. The discussions aimed to create an initial coding tree that can be used to classify the data, so that each code can be compared systematically with parts of the data set, as suggested by Gale et al. (2013). We found that we tended to agree on codes (how to label, which to merge, etc.) more quickly when the discussions consisted of at least three, instead of only two researchers. During points of divergence, a third or fourth person could weigh in on decision points about how to group the codes.

Stage 4: Developing a working analytical framework
Coding Phase Two: Question-based coding (Deductive)
After open coding, we met weekly to discuss the line-by-line codes and agree on overarching analytical and thematic codes. As we compared codes, we noted the benefit of open coding as it allowed the researchers to siphon through the text and minimized missing uncommon or unexpected information that could happen if only one researcher was reading the data. However, open coding is only an initial step and must be followed by a more structured or guided coding that could yield not just a familiarity of the transcripts but also an initial

description of the common and contrasting viewpoints of the participants. We started building a foundation for the coding tree by identifying common, uncommon, and unexpected codes.

Since the questionnaire was organized into eight blocks, we used question-based coding or block-based coding and prepared codes based on the interview guide questions before we highlighted and coded the segments from the transcripts. In question-based coding, we formatted the codes as follows: "block number_question" or "topic_answer_respondent type". Coding this way, we noticed how answers had different "directions" (positive, negative, or neutral). The contrasts across and within-participant groups became more apparent. This familiarization with the data later helped us decide that codes should carry different directions and could be applicable regardless of roles. The contrasts across stakeholders could be considered in the analysis using the transcript code titles instead of the codes themselves.

Using this approach halved the number of codes which were also easier to group (from the initial descriptive codes, to the categorical and analytical codes) because the coding is based on the types of answers as opposed to the initial descriptions in open coding where two independent researchers can code a similar line in different ways. Still, the codes were too many (around 150) and had too many levels (four) to be used for analysis. We added two researchers at this stage to workshop the initial versions of the analytical framework. At this point, the codes already had descriptions, forming the basis of the coding tree's outline.

Coding Phase Three: Feature-based coding (Combined)
The goal of the workshops was to create a working analytical framework that would facilitate the classification or coding of all the stories from the interview data, also known as indexing transcripts. In the workshops, we discussed the boundaries (inclusions and exclusions) of all the codes. The boundaries were important so the codes can be used in tagging (assigning into codes) the pertinent lines from the transcripts.

The main criteria for building and revising the analytical framework were relevance or usefulness to the data (i.e. they can describe the stories succinctly), and ease of use (i.e. the coders are not confused by what the codes pertain to, or how they would be applied to the transcripts). We aimed for coding tree that had a clear logical flow, but concise enough to cover all possible stories discussed in the interviews. The final coding tree or analytical framework consisted of seven themes branching out to 20 major codes, which in turn branched out to 121 minor codes.

Stage 5: Segmenting and reviewing the segmentation of transcripts
Segmenting establishes which parts of the transcripts were to be coded, creating a constant or a fixed part of the transcripts for indexing. We agreed that a segment should carry a complete thought, argument, or story. A segment is composed of premises or observations that lead to an argument, as we observed that it was easier to apply codes to more contextualized and thus more meaningful lines. Thus, our segmenting style leaned toward larger segments rather than splitting the transcripts into phrases. A segment begins with the interview prompt or question, and concludes with the end of a completed argument or thought, often containing subtopics in between. Segments need to include the questions so that the coders can contextualize the discussion, or know how or why the participant shared their observation. Using MS Word, we highlighted these complete thoughts or stories in alternating colors for easier coding. The subject of the argument is the story to be coded.

We conducted memoing to keep track of emerging ideas or analysis, such as when we sensed a possible explanation, pattern, or a hunch as we read and segmented the transcripts. It also helped in marking new research questions that can be explored further in later studies such as when there are stories not directly prompted or related to the questions in the interview guide and yet offered new insights on the experience of curriculum implementers. Memoing helps in reflecting on the process of using evidence to write our initial findings. For example, we referred to our memos to document our segmenting decisions.

Reviewing segmentation increased reliability by ensuring that the segmentation rules were applied. A good indicator of a well-segmented transcript is if the reviewer can easily apply the codes to the segments (i.e., each segment carried a completely developed or articulated narrative, argument, or thought). Reviewing was done asynchronously using the comment panel on MS Word where the reviewer could suggest expanding or contracting the coverage of highlights, such as if a reviewer finds that some lines still fall under a preceding code or otherwise, or a new segment needs to be made for a different code. This stage helps the segmenters solidify their indexing decisions as it also acts as peer consultation on possible ways of coding the segments.

Stage 6: Tagging or applying the analytical framework

At this stage, we apply the codes of the analytical framework or coding tree to the rest of the transcripts. The codes are abbreviated so they can be more easily applied (written) into the segmented transcripts. In our practice, the number of letters in the abbreviation corresponded to levels in the coding tree. Our coding tree had three levels corresponding to the theme, major code, and minor code. For example, "f_co_bal" represents "features_content_balancing cognitive demand". One letter pertains to the first level or the theme, two letters for the second level or the major code, and three letters for minor codes.

We used an Excel sheet detailing the coding tree's themes, major and minor codes, definitions, inclusion and exclusion criteria. The sheet could be used to do a word search of key topics included in a code's boundary. Aiming to make the coding tree easy to use to minimize errors in assigning the codes, we created mental shortcuts based on the themes' boundaries. For example, the first theme (feature) pertains to descriptions of the curriculum while the second and third themes (implementing and alternative curricula) pertains to user assessment, experience, and prospections and retrospections on the curriculum.

Stage 7: Charting data into the framework matrix

After tagging the segments from all transcripts using the comment panel on MS Word, the comments were extracted using the DocTool extension which collected all the comments containing the tags and the corresponding segments. We arranged the tagged segments into a table with columns for the codes, and rows for the representative statements. Then using Stata, the extracted comments were collated in an Excel sheet according to transcript code, code, representative statement, respondent, the school characteristics (i.e., urban/rural, class level offering, size, school/division, offering Indigenous Peoples education), block (related to the random sampling of the schools), treatment (pilot or nonpilot), segment number, theme, minor code, and major code. Once indexed, we could sort the data according to a code of interest and see all the coded representative statements (the segmenting of transcripts). Using the column filter options in Excel, we could view only the coded segments from the teachers' transcripts, or from the urban schools, depending on which perspectives we want to investigate.

Stage 8: Interpreting

The framework matrix could show the top (most tagged) and marginal (less tagged) codes in a scree graph. The most tagged codes could be extrapolated as the topics most discussed during the interviews. We thus began writing our report on the most tagged codes. Interpretation involved summarizing the similarities and differences in the perspectives within and across the three participant groups, and writing a synthesis of these perspectives taken together. The output is a list of areas of the curriculum that work and areas that need further support or fine-tuning.

The framework matrix records the audit trail of the codes analysis. Using Excel sheets in summarizing the coded segments with the representative statements, we document the basis our analysis or data interpretation. To surface the participants' multiple perspectives within and across groups, we formatted the comparisons using a 2 x 3 table with participants in rows, and the school characteristics in columns. Following the constant comparative method, we noted the similarities and differences within and across groups while presenting evidence or representative statements that illustrate the major and minor themes or findings. After organizing the representative statements and themes in this format, we then wrote the main stories (major themes or similarities in views), insights (strange, new, unseen stories), and suggestions or ways forward or strategies relevant to the concerns. With this method, researchers could move back and forth across the data until a coherent narrative emerges (Gale et al., 2013).

We considered the research questions (how to frame the data) to make decisions on how to make the stories cohere. We analyzed according to the school characteristics (urban, rural) per participant group, and then noted the rest of the school characteristics for every highlighted representative statement. The groupings of similarities and differences emerged as we combed through the coded segments and gradually composed the summary (commonalities across groups), insights (caveats or emerging themes), and strategies (suggestions for addressing the concerns). We noticed differences in the answers of the participants, such that teachers and principals usually discussed their experiences with more detail, while representatives from the administration often gave statements that refer to policy and views that sometimes conflicted with what principals and teachers narrated. Within and across groups, however, teachers and principals may also have divergent views. Following Collaco et al. (2021), differing perceptions were incorporated in the triadic summary. With these, we could identify caveats according to the respondent groups, questions that could be probed in continuing studies, and findings that could be validated using survey data.

For example, we used Fisher's exact test to check for significant differences on the information sources that control group teachers used to learn about the treatment curriculum. As shown in Table 3, of 27 non-pilot teacher FGD transcripts, 18 were tagged for learning about the pilot curriculum through in-service training. Four transcripts were from the control group and 14 from the treatment group. Learning about the treatment curriculum is significantly correlated with whether the teachers are from control and treatment groups. More teachers from the treatment group said they learned about the pilot curriculum through in-service training (Fisher's exact test = 0.011). This example shows that doing quantitative analysis using qualitative data can be done using the framework method.

Table 3. Association test for source of information and treatment group

Learned about pilot curriculum through in-service training?	Control school teachers	Treatment school teachers	Total
No	7	2	9
Yes	4	14	18
Total	11	16	27

5. Modifying the framework method: increasing consensus

We modified Gale et al.'s (2013) stages such that our approach included eight stages, with the modification mostly in stage 4 or developing the working analytical framework. Essentially, our modification aimed to organize evidence systematically, ensuring that "analysis transcends the imagination of a single individual" (O'Connor and Joffe 2020). After multiple checkpoints, we found that the key to getting high intercoder reliability was segmenting the transcripts, which fixes the boundaries of the segments, so we have a constant framing of the data. Otherwise, having each researcher segment the transcripts idiosyncratically resulted in lower intercoder reliability. Thus, we added stages for segmenting and reviewing segments to the FM stages. O'Connor and Joffe (2020) noted that segmenting is amenable to intercoder reliability but at the expense of analytical flexibility. We chose to proceed with segmenting as we prioritized interpretive consensus and high reliability given the large-scale triadic data to be analyzed for the assessment.

Increasing consensus using intercoder reliability checkpoints

To get ICRs, the segmented transcripts were tagged or indexed in the appropriate code using a word processor. Then, we extracted the codes for each segment using DocTools. Those were compiled using MS Excel and Stata to create variables indicating whether each code was tagged by each rater for every transcript. With Stata, the tags were formatted so ICR could be computed using Fleiss' Kappa and Pearson's correlation coefficient to check the similarities of tags done by the team members. The indicator variables for each rater (1 = tagged in transcript; 0 = not tagged in transcript) were then used to compute the Kappa statistic for more than two raters, as formulated by Fleiss et al. (2003). Pearson's pairwise correlation coefficient (Pearson, 1896; Pearson and Filon, 1898) was also used to identify which pairs of members have higher or lower levels of agreement.

We checked for intercoder reliability multiple times using five transcripts. These checkpoints were done as we constructed the analytical framework. The ICR measures how similarly we understood the analytical framework and applied the coding guidelines. A low ICR score means lower agreement about the codes' definitions and application practice. A higher ICR score means that we agreed on what each code meant and when to apply it. The interpretation of the Fleiss Kappa scores is indicated in Table 5. As shown in Figure 1, in initial checkpoints 1a and 1b we had a fair intercoder reliability indicating that we did not have a uniform application of the codes possibly because of varying interpretations of the definitions per code. This might also be because some codes were related or similar resulting in different codes assignment or tagging.

To address the low ICRs, we conducted workshops to discuss how to group, label, and refine the codes' definitions, or set the boundaries of the codes' definitions. We identified which codes were used differently and focused our discussions on those. We refined the definitions by identifying areas of disagreement (different understandings) and finding consensus on those codes. Creating an exclusion and inclusion criteria (key words or key concerns are grouped under each code) helped to increase the ICR from moderate to substantial.

The codes should be easy to use such that judgement when tagging should follow clearly defined rules as to when to apply them. To minimize ambivalence, we agreed not to extrapolate from other possible definitions of the codes, and instead use the codes strictly as defined, adding to the inclusion and exclusion boundaries if needed. A concise coding tree with more broadly defined codes was easier to remember and apply than a more granular coding tree.

In checkpoints without segmenting, our intercoder reliability was only substantial even when using a coding tree with a more defined usage guideline, and clearly articulated logic. After numerous intercoder tests, we confirmed that the cause of the low agreement was because each coder would choose to highlight or segment transcripts differently before coding. A varied segmentation of the transcripts resulted in differences in the tagging or coding of transcripts. While we already had a substantial ICR, it was important to have an almost perfect ICR since we had hundreds of transcripts to code. We only trusted that the working analytical framework was comprehensive enough, i.e. had codes that could accommodate possible stories in the subsequent transcripts when we reached an almost perfect ICR. Once an almost perfect ICR was reached, however, the analytical framework remained iterative i.e., continually refined but mainly by adding new stories in the inclusion boundaries if needed when tagging the rest of the transcripts. Segmenting increased intercoder reliability as shown in checkpoints for transcripts 1 and 2. Only transcript 5 had a lower ICR possibly because of the four-month gap between the checkpoints a and b, coded in Sept 2024 and Jan 2025 respectively. Still, ICR for a and b are interpreted as almost perfect.

Intracoder reliability tests validate the effect of segmenting transcripts. It checks the reliability of each coder's usage of the coding tree by measuring any significant changes in how they apply the codes. To get the intracoder reliability, we tagged the checkpoint transcripts twice: once in September 2024, and another in January 2025. As shown in Figure 2, we had moderate to almost perfect intracoder reliability, prompting the need for segmenting to increase our intercoder consensus.

The codes' performance could also be checked using Fleiss Kappa, which we used to check whether the coders tagged each code across the transcripts we used for our checkpoints. By evaluating the code performance, we could identify which codes were tagged consistently and ensure those were included in the analysis, as we had higher interpretive consensus on those codes. In our ICR checkpoints, all coders tagged about a third of the codes, demonstrating high consensus on those codes. We ensured that the high-performing codes were included in the final report of the assessment.

Table 4. Method process and resources

Stages	Process	Time required	No. of researchers
Stage 1: Transcription	▪ Verbatim transcription of audio recording; Translation ▪ Data management and anonymization	5-8 hours per transcript of 1.5 interview hours	At least one per two schools
Stage 2: Familiarization with the transcripts	▪ Open coding using sample transcripts ▪ Coding via inductive approach by highlighting important phrases and creating initial labels or descriptive codes	1.5 – 3 hours per transcript	2
Stage 3: Coding	▪ Using Question-based coding or using deductive approach ▪ Highlighting transcripts and creating codes with more focus on the responses to the questions	1.5 – 3 hours per transcript	3
Stage 4: Developing a working analytical framework	▪ Merging codes to create categorical codes ▪ Clearly defining codes: Setting inclusion and exclusion boundaries ▪ Getting intercoder reliability (conducting checkpoints) ▪ Discussing codes with low agreement	1 to 2 days of weekly workshops for 1.5 months	5
Stage 5: Segmenting and reviewing transcripts	▪ Random assignment of transcripts to the coders ▪ Highlighting segments as guided by the logic of the coding tree ▪ Reviewing the segmentation of the transcripts ▪ Improving the ease of applying the available codes	1 to 2 hours per transcript	4
Stage 6: Tagging or applying the coding frame	▪ Applying the codes or indexing the segments	1 to 4 hours per transcript	3
Stage 7: Charting data into the framework matrix	▪ Extracting the tagged or coded segments using DocTools ▪ Creating a framework matrix using Stata	4 hours for all the transcripts	3
Stage 8: Interpreting data	▪ Thematic analysis through constant comparative approach ▪ Content analysis using tag frequencies ▪ Triadic analysis/Triangulation ▪ Writing the synthesis, insights, and strategies for addressing issues raised	2 days per highly tagged codes	3

Figure 1. Checkpoint for intercoder reliability

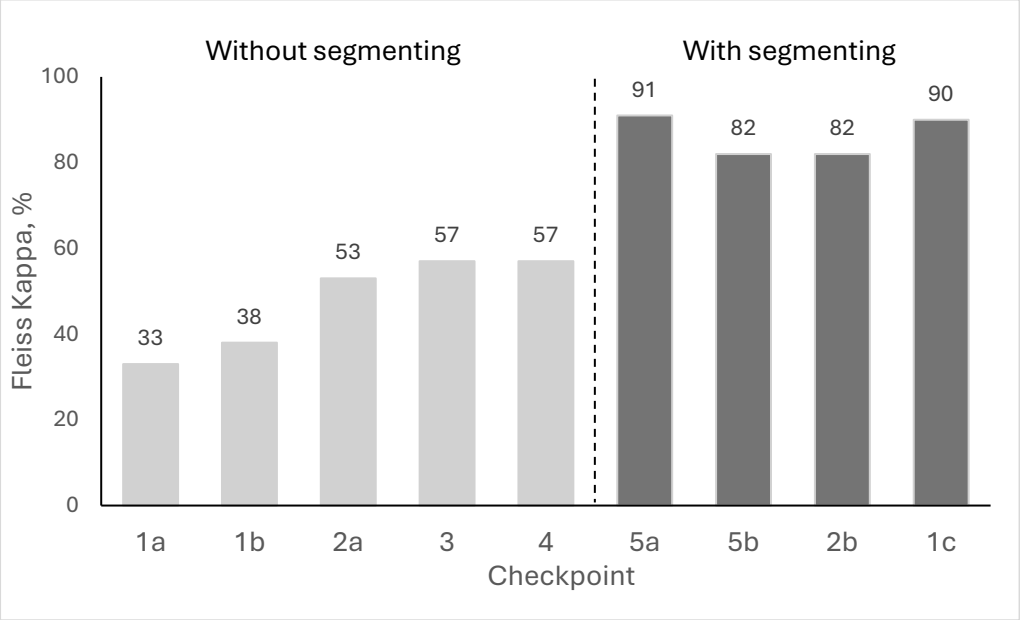

Figure 2. Checkpoints for intracoder reliability

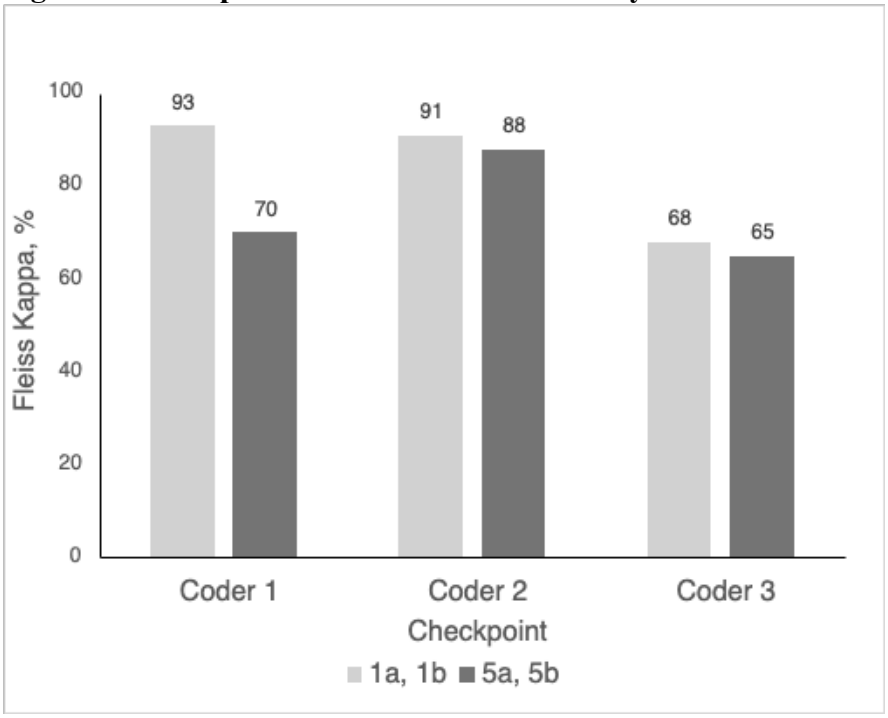

Table 51. Interpreting Fleiss Kappa

Fleiss Kappa	Interpretation
<0	Poor
0–20	Slight
20–40	Fair
40–60	Moderate
60–80	Substantial
80–100	Almost perfect

Figure 3. Snapshot of memoing on Excel sheet

Transcript no.	Seg no. and code	Notes	Memo category (optional)
S28NT	General	Ang haba ng transcript, it took me one whole afternoon to code this , meron siyang 87 codes	ON
S35PT	043_t_wb_sal	Salary should increase because they usually shoulder whatever amount the MOOE does not cover; hospitality practices taken on by teachers	TN
S25NT	034_i_te_int	Yung integrating content and spiral approach ay very magkapatid yan sila.. so can be confusing pero mas integrating ko to nilagay kasi ang kwento yung missed topics na nangyayari dahil sa integrative na approach or style ng curr	TN
S33NP	044_s_re_res	Note: this School Monitoring and Evaluation and Adjustment is not a test for students but it is more of managerial tool for the admin/principal; but I tagged this is because it is about using results of standard measurement tool (not an exam, but an M&E); we might need a new code for this actually, yung M&E tools na gamit ng admin	TN
S33NP	050_m_in_res	The problem with successful MTB-MLE is still the learning resources (orthography etc) -- how to implement it if wala tayo nung mga books for our languages; it seems hindi sapat yung organic na capacity ng teachers and learners since may imposition or select language na rin na pwede gamitin in the classrooms	TN

Increasing consensus using memoing throughout the stages

We practiced memoing throughout the stages of the framework method as it was useful in noting how we were making sense of the data, and in planning or strategizing our methodology as we reflected on challenges. In memoing, we are describing the data but also moving toward analysis. As such, it was helpful in the iterative construction of the analytical framework because we could park areas that need to be explored in-depth.

Memos were separated from the data and were used to note ideas that might be useful for later analysis. We wrote memos as we combed through the raw data and made tentative connections or linkages of how the stories fit together (or not). Using an Excel sheet, we noted the thought process or reasons for applying certain codes instead of others, or the questions to tackle with the research team during workshops. As shown in Figure 3, in the fourth column of a memoing sheet, the memos could be categorized as observation, methodological, theoretical, and personal notes (Gibbs 2012).

In combing through the transcripts, we encountered new stories not yet included in the codes' boundaries. Regular workshops were useful in catching these stories and collectively deciding how to classify them. In our practice, all the minor codes included an "other" code that acts as "nets" for insights from stories that do not fit the codes' boundaries. When "other" codes were applied, we recorded the reasoning in our memos. In the last stage, we return to the segments

tagged under the “other” codes. They provided caveats to the main findings, such as alternative perspectives, surprising (unpredictable), or fresh angles.

Memoing is a means of meeting COREQ’s standard of reflexivity in qualitative method, as it can be used as a paper trail of how the analysis developed over the course of the framework method. While the framework method is useful in organizing large-scale interview data for thematic analysis, there is still a risk of missing the context-and-text dynamic, i.e. reading between the lines, paying attention to the unsaid, etc., that is characteristic of in-depth QDA or line-by-line coding that can be done with fewer transcripts. Since the study handled hundreds of interviews, memoing helped minimize the tendency for mechanical or face-value reading or analysis.

6. Conclusion

The framework method can be applied in research contexts that require both quantitative and qualitative evidence, in contexts where data are more qualitative, or when questions require qualitative explanations. By synthesizing the interviewees' stories using codes, FM can be used for content analysis through code frequencies. Researchers can conduct the constant comparative approach associated with thematic analysis by displaying evidence per code and participant group. Displaying the qualitative data in these ways can help substantiate survey findings. In addition, researchers using FM can combine the inductive and deductive approaches of qualitative data analysis for large-scale data, addressing predefined questions while also surfacing new research problems from the analysis through consistent memoing.

In creating a working analytical framework, it is useful to aim for a logical summary of the stories found in the interviews (create a story structure) that can be easily understood by readers (such as coders, participants, or study audiences). To use the coding tree, researchers need checkpoints ensuring intercoder and intracoder reliability (i.e., consistency of judgement among, and for each coder). This way, the method can be used by a team of diverse researchers, enabling them to divide the analysis of large-scale interview data while reaching a consensus on how to interpret and build evidence from data.

References

Abrigo, M. R. M., Lingatong, E. L., Paquibot, J. L., Daga, E. D. C., Perez, T. D. S., Relos, C. M. L., Apostol, G. C. Forthcoming. Monitoring and analyzing correlates of basic education outcomes at the school level: Assessment of the Department of Education’s MATATAG K to 10 curriculum pilot implementation.

Arifin, S. R. M., Cheyne, H., Maxwell, M. and Pien, L. S. 2019. Framework analysis: A worked example from midwifery research. *Internation Journal of Quality in Healthcare, 9*(6), 349-357. DOI: 10.1093/intqhc/mzm042.

Collaco, N., Wagland, R., Alexis, O., Gavin, A., Glaser, A., and Watson, E. 2021. Using Framework method for the analysis of qualitative dyadic data in health research. *Qualitative Health Research*, 31(8), 1555-1564. https://doi.org/10.1177/10497323211011599

Department of Education. 2024. *DepEd's MATATAG Agenda, BEDP 2030 approved by NEDA Board as National Policy and Plan for Basic Education.* Retrieved January 23, 2025 (https://www.deped.gov.ph/2024/04/27/depeds-matatag-agenda-bedp-2030-approved-by-neda-board-as-national-policy-and-plan-for-basic-education/).

Fleiss, J. L., Levin, B., and Paik, M. C. 2003. *Statistical Methods for Rates and Proportions*. 3rd ed. New York: Wiley.

Gale, N.K., Heath, G., Cameron, E. Sabina, R. and Redwood, S. 2013. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC Med Res Methodol* 13, 117. https://doi.org/10.1186/1471-2288-13-117

Gibbs, G. R. 2012. *Writing memos for qualitative research* [Video]. YouTube. https://www.youtube.com/watch?v=eXj4QiyZl50&t=1470s

Green, K. W., Hughes, E. Small, R. Welch, and N. Gibbs, L. Daly, J. 2007. Generating best evidence from qualitative research: the role of data analysis. A*ustralian and New Zealand Journal of Public Health*, 31 (6), 545-550. https://doi.org/10.1111/j.1753-6405.2007.00141.x. https://www.sciencedirect.com/science/article/pii/S1326020023014371

Hesse-Biber, S. N. 2017. *The practice of qualitative research.* Sage Publications. Inc.

Klingberg, S., Stalmeijer, R. E., and Varpio, L. 2024. Using framework analysis methods for qualitative research: AMEE Guide No. 164. Medical teacher, 46(5), 603–610. https://doi.org/10.1080/0142159X.2023.2259073

O'Connor, C. and Joffe, H. 2020. Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines. *International Journal of Qualitative Methods,* 19. https://doi.org./10.1177/1609406919899220.

Pearson, K. 1896. VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London Series a Containing Papers of a Mathematical or Physical Character*, 187, 253–318. https://doi.org/10.1098/rsta.1896.0007

Pearson, K., and Filon, L. N. G. 1898. Mathematical contributions to the theory of evolution. IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Proceedings of the Royal Society of London*, 62(379–387), 173–176. https://doi.org/10.1098/rspl.1897.0091

Ritchie, J., Lewis, J., Nicholls, C. M., and Ormston, R. 2003. *Qualitative research practice: A guide for social science students and researchers.* London: Sage.

Tong, A., Sainsbury, P. and 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