

CHAPTER FOUR: CONTENT AND THEMATIC ANALYSIS

Hélène Joffe and Lucy Yardley

Aims of this chapter

- i) To introduce the basic principles of content and thematic analysis
- ii) To explain in detail how to code qualitative data
- iii) To consider the role of computer software in qualitative data analysis
- iv) To discuss the advantages and limitations of content and thematic analysis

Content analysis is the accepted method of investigating texts, particularly in mass communications research. Most content analysis results in a numerical description of features of a given text, or series of images. **Thematic analysis** is similar to content analysis, but pays greater attention to the qualitative aspects of the material analysed. This chapter considers the thinking that lies behind each of the two methods, as well as showing how content analysis and thematic analysis are conducted. The roles played by theory, coding and computer packages in such analyses are highlighted.

Introduction to content analysis and thematic analysis

Content analysis involves establishing categories and then counting the number of instances in which they are used in a text or image. It is a partially quantitative method, which determines the frequencies of the occurrence of particular categories. However, while early proponents (e.g. Berelson, 1952) conceptualised content analysis solely in terms of counting the attributes in data (e.g. words), more recent writings on content analysis (e.g. Krippendorff, 1980; Bauer, 2000) contain a broader vision. Krippendorff's point of departure is that social scientists tend to regard data as symbolic phenomena, and since symbolic data can always be looked at from different perspectives, the claim that one is analysing *the* content is untenable. In other words, talk about, for example, 'distress' cannot be taken as a straightforward observation of the phenomenon distress in the same way that a measurement of heart rate may be taken as a direct observation of one aspect of heart functioning. This is because talk about distress has much more complex symbolic relationship to feelings of distress than heart rate does to cardiac functioning. Thus while "content analysis is a research technique for making replicable and valid inferences from data to their context" (Krippendorff, 1980:21), messages do not have a single meaning waiting to be unwrapped. The person analysing communications always has to make inferences, but these should be made by systematically and objectively identifying characteristics of the text.

The content analytic method is appealing because it offers a model for systematic qualitative analysis with clear procedures for checking the quality of the analysis conducted. However, the results that are generated have been judged as 'trite' (Silverman, 1993) when they rely exclusively on frequency outcomes. Researchers employing this method are also sometimes accused of removing meaning from its context. The problem is that a word or coding category such as 'pain' may occur more

frequently in the talk of one person or group of people than another for many reasons; frequent occurrence could indicate greater pain, but might simply reflect greater willingness or ability to talk at length about the topic, or might even occur in repeated assertions that pain was not a concern. Thematic analysis comes into its own in terms of these two criticisms. Ideally, it is able to offer the systematic element characteristic of content analysis, but also permits the researcher to combine analysis of the frequency of codes with analysis of their meaning in context, thus adding the advantages of the subtlety and complexity of a truly qualitative analysis.

Thematic analysis shares many of the principles and procedures of content analysis; indeed, in Boyatzis' (1998) conceptualisation of thematic analysis the terms '**code**' and '**theme**' are used interchangeably. A theme refers to a specific pattern found in the data one is interested in. In thematic and content analysis, a theme of a coding category can refer to the **manifest** content of the data, i.e. something directly observable, such as mention of the term 'stigma' in a series of transcripts. Alternatively, it may refer to a more **latent** level, such as talk in which stigma is implicitly referred to (e.g. by comments about not wanting other people to know about an attack of panic or epilepsy). Thematic analyses often draw on both types of 'theme', and even when the manifest theme is the focus, the aim is to understand the latent meaning of the manifest themes observable within the data, which requires interpretation.

A further distinction in term of what constitutes a 'theme' (or coding category) lies in whether it is drawn from existing theoretical ideas that the researcher brings to the data (**deductive coding**) or from the raw information itself (**inductive coding**). Theoretically derived themes allow the researcher to replicate, extend, or refute prior discoveries (Boyatzis, 1998). For example, the researcher might code patients' talk about a treatment programme that they were about to follow using coding categories based on the elements of the Health Belief Model or the Theory of Planned Behaviour, in order to determine to what extent each model seemed to capture their spontaneously expressed attitudes and beliefs. However, more inductive themes, drawn from the data, are often useful in new areas of research (although it should be noted that no theme can be entirely inductive or data driven, since the researcher's knowledge and preconceptions will inevitably influence the identification of themes). A key dilemma facing the analyst is whether to 'test' theory, or to explore new links. For example, if a theme corresponding to an expressed belief in an African geographic origin for an epidemic emerges prominently in two consecutive studies looking at western respondents' ideas about epidemics, is it an inductive theme in the first and deductive in the second, or inductive in both? In addition, such a theme refers to manifest content (i.e. the African origin of the disease), but also conveys strong latent meaning when it is westerners who consistently utter it (i.e. attribution of the origin of the disease to a distant place and alien culture).

In the type of thematic analysis proposed in this chapter, existing theories drive the questions one asks and one's understanding of the answers, so that one does not 'reinvent the wheel'. This is important since qualitative work, to a greater degree than quantitative research, has the potential to underplay evidence that contradicts the assumptions of the researcher. Therefore, it is advantageous to hold a model of 'testing' in mind, regarding taking counter-evidence seriously, even though it is only in quantitative work that the researcher 'tests' theories in a statistical sense.

Conducting content and thematic analyses

What to code

In clinical and health psychology, transcripts of interviews often form the data upon which the content or thematic analysis is conducted, although open-ended answers to a questionnaire, essays, media and videotaped materials can also be used (Smith, 2000). While the steps involved in conducting a content analysis are well established (see Bauer for a review of the different types of content analytic design and how to sample material for a content analysis), there are surprisingly few published guides to how to carry out thematic analysis, and it is often used in published studies without clear report of the specific techniques that were employed. These techniques will therefore be detailed and evaluated below.

For both content and thematic analysis there is a need to create conceptual tools to classify and understand the phenomenon under study:

“This involves abstracting from the immense detail and complexity of our data those features which are most salient for our purpose.” (Dey, 1993:94)

This is done by way of coding, which is the widely accepted term for categorising data: taking chunks of text and labelling them as falling into certain categories, in a way that allows for later retrieval and analysis of the data.

What one chooses to code depends upon the purpose of the study. Bauer (2000) warns against adopting a purely inductive approach where one codes whatever one observes in the text. Rather, codes need to flow from the principles that underpin the research, and the specific questions one seeks to answer. The total set of codes in a given piece of research comprises the **coding frame** (a term used interchangeably with ‘coding manual’ or ‘coding book’). Such a frame is given coherence by being derived from higher-order ideas, and Bauer’s argument is that codes should be derived from existing theory. However, there would be little point in doing research if one were not simultaneously open to the data and what it might offer anew in terms of the theory’s development or refutation. The point of the coding frame is that it sets up the potential for a systematic comparison between the set of texts one is analysing (Bauer, 2000); it is by means of this frame that one is able to ask questions of the data (see below).

How to code

Coding in content and thematic analytic research is taxing and time-consuming because there are generally no standardised categories. The researcher codes in order to answer the research questions, and the coding frame is developed in a manner that allows for this. Coding involves noting patterns in the data and dividing up the data to give greater clarity regarding its detailed content. In order to do this, the patterns are labelled with codes. Distinctions are drawn between different aspects of the content by organising the data into a set of categories.

An early decision must be taken regarding what the unit of coding will be, i.e. whether codes will be attached to each line of text, sentence, speaker turn, interview or media article. For example, one might simply want to know how many interviewees or newspaper articles mentioned ‘stress’, in which case the coding unit would be the entire interview or article. However, generally coding is much more fine-grained, using a coding unit such as the sentence or a phrase, which allows the researcher to count how often in a single interview or article the code occurs, and to analyse the relationship of this code to other codes, in terms of co-occurrence or sequencing. For example, one could then analyse whether mention of ‘stress’ was typically associated with mention of ‘work’, and whether talk about work was typically followed by talk about stress, (which might be an indication that work was

viewed as leading to stress) or whether talk about stress usually preceded talk about work (which might reflect talk about the impact of stress on work).

A related issue that the researcher must resolve is whether each coding unit must be coded exclusively into just one category, or can be coded into multiple categories. The larger the coding unit, the more likely it is that it will contain material that could be coded into more than one category. Note that for some quantitative analyses of codes the codes must be **exclusive**, since statistical analysis of the relationship between codes will assume this (e.g. it is not possible to calculate the correlation between two codes if they are not independent -- in other words, if a single text segment can be assigned both codes). Exclusive coding has the advantage that it forces the researcher to develop very clearly defined coding categories, and this can enhance the development of the theoretical basis for these coding decisions. For example, if one wishes to code exclusively using the concepts of 'coping strategies' and 'handicap (i.e. changes in lifestyle due to illness)', then the task of deciding which text segments should be assigned which code will oblige the researcher to think very carefully about the definition of and relationship between these concepts; if someone with dizziness says that they were so dizzy that they 'had' to hold on to a railing, should this be coded as an instance of coping (by holding on) or handicap (as they were unable to walk unaided) or both? In a more qualitative analysis the researcher may reject the necessity of making such a distinction, which may appear arbitrary and artificial, and will instead allow the codes to overlap on occasion. However, it is important that codes should not be too broad and overlapping, or they will not serve the intended purpose of making distinctions between different aspects of the content, and the researcher will find that he or she is categorising large chunks of the data using the same multiple set of codes.

Insert Figure 1 about here

A code should have a label, an operationalisation of what the theme concerns and an example of a chunk of text that should be coded as fitting into this category. For example, the range of codes that arose from asking a sample of Zambian adolescents to talk about the origin of AIDS is shown in Figure 1. This example shows how coding categories very often form a hierarchy, with a small number of higher level categories (such as 'origin' in this example) that can be progressively sub-divided into lower level sub-categories (in this example the first level of sub-category is 'geographical' vs. 'God' or 'practice', and the lowest level of sub-category corresponds to the continents within the sub-category 'geographical' and the types of behaviour under the sub-category 'practice'). Sometimes this hierarchy is created by coding the text first at the highest level, and then developing finer coding discriminations that can be used to create sub-categories that fall lower down in the hierarchy. For example, if using predefined theoretical high level categories for deductive coding, such as the elements of the theory of planned behaviour, the researcher might first code the text into material relating to attitudes, subjective norms, and perceived control, and then inductively construct coding categories which distinguish between different sub-categories of attitudes, norms and control. However, the researcher might decide that two or more sub-categories within a code are so distinct that two entirely new, separate codes need to be created by **splitting** the original coding category.

Alternatively, the researcher may begin by using very specific low-level codes when coding at the initial textual level (as in grounded theory analysis, see Ch 5). For a coherent analysis, fewer, more powerful categories will usually be required and so the initial textual categories are integrated into conceptual categories by way of ‘**splicing**’ and ‘**linking**’, according to Dey (1993). Splicing can be thought of as the opposite of dividing up material. It is the fusing together of a set of codes under an overarching category. It involves increasingly focusing the categorisation activities in the knowledge that it will be impossible to incorporate all codes into the final analysis. This process of fusion involves the researcher thinking through what codes can be grouped together into more powerful codes. For example, the codes European, US and Australian origin, in the coding frame above, could be fused into a ‘Western origin’ code. An alternative to actually splicing codes is to create links between codes. Multiple themes can be clustered into groups – the particular cluster or link is a higher order theme allowing for higher order abstraction and interpretation. One might do such clustering to show a conceptual relatedness of themes, or to show the sequential relatedness of sets of themes across the data. Thus, if the codes European, US and Australian origin were clustered rather than spliced, it would be possible to analyse them jointly, to explore themes which were common to Western origin countries, and also separately, to determine whether there were themes unique to particular countries. The splicing and linking processes can be driven by theoretical concerns, policy issues or purely grounded in the data itself.

Moving from coding to analysis

Once the codes have been developed, refined, and clearly described in the coding manual, the researcher may determine the reliability with which the codes can be applied. An initial impression of reliability can be gained by applying the codes to the same piece of text on two occasions separated by a week or so (a kind of ‘test-retest’ reliability). Although the coding will be influenced by similar subjective processes on both occasions, consistent coding by the researcher at least indicates that the distinctions made between codes are clear in the researcher’s mind – if you cannot apply the codes consistently, there is no possibility that anyone else will be able to! The stronger test of reliability is to calculate the correspondence between the way in which codes are assigned to the text by two independent coders (see chapter 7 for details of methods of calculating **inter-rater reliability**). If you wish to claim that your codes are objective, reliable indicators of the content of the text then you must demonstrate that the inter-rater reliability of your coding frame is good. Reliability testing is commonly used in content analytic work, especially if quantitative analysis is to be employed.

Inter-rater reliability checks are not always used in thematic analytic research since there is scepticism regarding such tests: it can be argued that one researcher merely trains another to think as s/he does when looking at a fragment of text, and so the reliability check does not establish that the codes are objective but merely that two people can apply the same subjective perspective to the text. However, this criticism overlooks the value of having to make the interpretations of the data very explicit and specific in order to achieve reliable coding. The more clearly the rationale for coding decisions is explained in the coding manual, the higher the inter-rater reliability will be; for example, it may be helpful to detail the logic underlying subtle coding discriminations (which may have been debated when disagreements arose between two coders), or to provide ‘negative’ examples of text segments which might appear to belong to that coding category but actually do not. When carrying out a complex thematic analysis, inter-rater comparisons provide a valuable opportunity to open up

the rationale for the coding frame to the scrutiny of others, to examine and discuss the reasons for any differences in coding decisions, and hence to fine-tune the theoretical bases and operational definitions for the coding categories. This process not only means that the second researcher will code most of the transcript in a similar way to the first researcher, but also that other researchers looking at the system of coding used will find one that is fairly transparent, coherent and understandable, as opposed to an idiosyncratic, opaque system of interpretation devised by a single researcher.

Insert Box 1 about here

When all of the data have been categorised, the analysis can begin. A code can be used in a primarily quantitative manner, in which the numbers assigned to codes form the basis for a statistical analysis. Depending upon the way in which codes have been applied to the data (see Boyatzis pp.141/2), statistically based analyses such as correlations, group comparisons (e.g. using the chi-square test), cluster analysis or even multiple regression can be conducted. This would be the more usual route for a content, rather than a thematic, analysis. On the other hand, codes can be used for a purely qualitative analysis, where the focus tends to be on description of verbal patterns. Perhaps it is a point between these two positions that is most appropriate for a thematic analysis: the nuances of the high frequency themes are explored in depth. This approach is particularly fitting for research that is driven by social representations tenets; themes widely shared within particular groups are taken to illustrate the existence of social representations (see Joffe and Haarhoff, 2002).

Huberman & Miles (1994) spell out a useful sequence that summarises the method of generating meaning from a set of transcripts. Patterns and themes are noted; themes are counted to discern whether a theme is common or more rare; a plausible story is extracted from the data that can be related to the literature review; differences in terms of gender, class or other groupings are looked at; disconfirming as well as confirming evidence is examined. In line with a loose definition of 'testing' a research question, one looks at whether the original hunch can be sustained or not, and whether it needs modifying, as well as at the direction in which further research might go. New insights can often be provoked by attempting to understand what appear to be anomalies.

Computer Assisted Qualitative Data Analysis Systems (CAQDAS)

Over the past decades a series of computer packages (e.g. Ethnograph, Atlas ti, NUDIST), which aid content and thematic analyses have been produced. Their role in the process of analysis, advantages and disadvantages require careful consideration.

The central analytic task of thematic analysis, in particular, is to understand the meaning of texts. This requires researchers' minds to interpret the material. The computer is a mechanical aid in this process. Computers cannot analyse textual data in the way that they can numerical data. Yet, as a mechanical aid, the computer is able to make possible higher quality research for the following reasons. CAQDAS allow researchers to deal with many more interviews than manual analyses can.

Consequently, useful comparisons between groups can be made due to the inclusion of large enough numbers of participants in each group. The researcher is also assisted to look at patterns of codes, links between codes and co-occurrences in a highly systematic fashion, since retrieval of data grouped by codes is made far easier.

CAQDAS permit retrieval of data combinations in a manner similar to literature search computer packages, typically using combination retrieval terms such as AND, OR and so on. The researcher can therefore instantly retrieve, for example, all text

segments from older female interviewees which were categorised as relevant to 'perceived control', and compare these with similarly coded text segments in interviews with younger women, or with older men. If this process is carried out using, for example, cut-and-paste techniques (using either actual printed transcripts or a word-processing package), it is not only much more time-consuming, but literally cuts the text segment out of the (con)text, whereas the packages also allow the researcher to easily view the context of a particular coded text segment, so that the contextual meaning is not lost. Packages such as Atlas ti allow researchers to examine the patterning of themes across the range of interviews, and the common pathways or chains of association within interviews. More specifically, the filtering functions of packages such as Atlas ti allow the researchers to retrieve the patterns of codes prevalent in particular demographic groups, and such patterns can be retrieved as frequency charts, lists of textual excerpts, or visually, as visual networks.

Insert Figure 2 about here

Figure 2 provides an example of a visual hierarchy generated in Atlas ti, based upon the codes shown in figure 1 (with the European, US and Australian geographical origin codes fused into a 'West' origin code), which shows how such a package can usefully illustrate the findings of a thematic analysis. The higher up the code is in the table, the greater the number of people who mentioned it spontaneously. This figure makes it possible to see at a glance, for example, that a Western origin of AIDS is associated, in the minds of Zambian adolescents, chiefly with the practices of bestiality and science, which are viewed as linked, and to a lesser extent with anal sexual practices and inter-racial mixing. A diagram summarising the views of westerners would be very visibly different to this (see Joffe, 1999).

In summary, CAQDAS can provide an efficient way of retrieving text segments for systematic comparison, enumerating the degree of empirical support for different themes, and mapping the relationships between themes. It cannot fulfil the central task of textual analysis - to decode the meaning of the text - but as a mechanical aid to managing material it can facilitate it. Not only can it allow development of increased complexity of thought, since it can store and retrieve many more links than researchers can store in their minds, it also helps the researcher to assess how much counter-evidence exists for alternative interpretations. This is important, as qualitative work, in particular, has been accused of failing to take heed of trends that run counter to those that the research highlights. If the researcher is using qualitative and quantitative methods in tandem, the package can convert codes into frequencies and transfer them to SPSS for quantitative analysis, provided that the categorisation process meets the criteria for a more quantitative analysis (see above). When used in a thoughtful way, computer packages allow one to be highly systematic in a manner that is faithful to the data.

Evaluation of content and thematic analysis

Hollway and Jefferson (2000) challenge the practice of coding data into fragmented text segments in order to make sense of it. They claim that qualitative data analysis is one of the most subtle and intuitive human epistemological enterprises, and is therefore one of the last that will achieve satisfactory computerisation. They also state that fragmentation results in neglect of the whole, whereas the whole interview is not only greater than the sum of its parts, but by 'immersion' in the whole, one gains understanding of the parts, rather than vice versa. Their work, in line with the

narrative psychological tradition, consequently stays with people's storyline as a whole. Their critique is similar to the more general critique of thematic analysis, that it abstracts issues from the way that they appear in life, organising material according to the researcher's sense of how it connects, rather than the inter-relationship of themes in the participant's mind or lifeworld (see Boyatzis). However, it can be argued that the goals of thematic and content analysis are simply different from those of, for example, narrative analysis. The aim is to describe how thematic contents are elaborated by groups of participants, and to identify meanings that are valid across many participants, rather than to undertake an in-depth analysis of the inter-connections between meanings within one particular narrative

Contemporary uses of content analysis reveal a rather rich interpretation of the method; systematic research need not be sacrificed in the name of a complex unravelling of data. This mixture of systematicity and complexity appears to be an aspiration of content and thematic analysis alike. However, there are cases in which one or other should be used. Clearly, content analysis should be used if the aim is to carry out quantitative analysis. However, on a small sample size only the descriptive use of thematic coding is advisable, since it is meaningless to assign numbers to a data set that is too small to meet the usual minimum requirements for statistical analysis. For instance, if the sample cannot be regarded as large enough to permit reliable statistical generalisation to the population from which it was derived (see Ch?) then it may be misleading to report the frequencies of codes, as this would seem to imply that the frequency of occurrence in the sample was in some way representative of the likely frequency of occurrence in the wider population. Nonetheless, it can be helpful to give some indication of whether themes occurred rarely or commonly, although this can of course be conveyed using qualitative terms such as 'most', 'some' or 'a few'.

A good content or thematic analysis must describe the bulk of the data (Dey, 1993), and must not simply select examples of text segments that support the arguments one wants to make (see Silverman, 1993). Moreover, even if one quantifies the text for purposes of analysis, the analysis remains partially qualitative. In other words, it is vital to remember that numbers do not tell the whole story - that the number of times a category appears does not necessarily indicate the extent to which it is relevant to interviewees. A point that is only mentioned once, by one person, can still have great empirical relevance and conceptual importance. The aspiration of thematic analysis, in particular, is to stay true to the raw data, and its meaning within a particular context of thoughts, rather than attaching too much importance to the frequency of codes which have been abstracted from their context.

Insert Box 2 about here

Summary

Content and thematic analysis share the potential to be systematic, to rely on naturally occurring raw data, and to deal with large quantities of data. Thematic analysis, conducted in the way laid out in this chapter, allows for systematic analysis of the meanings made of the phenomena under investigation. Moreover, it is a form of analysis that is acceptable and meaningful to both researchers who normally employ quantitative methods and those who prefer a qualitative approach (Boyatzis, 1998). However:

“The challenge to the qualitative researcher is to use thematic analysis to draw the richness of the themes from the raw information without reducing the insights to a trivial level for the sake of consistency of judgement.” (p.14).

The chapter has demonstrated that computer packages, counting and the theory-driven nature of questions and means of answering them can contribute to the quality of work on naturally occurring texts (and indeed images). When used in a considered fashion, content, and particularly thematic analysis, allow the researcher to be faithful to the data while producing high quality social science.

Recommended reading

Bauer, M. W. (2000). Classical content analysis: a review. In M. W. Bauer and G. Gaskell (eds) *Qualitative researching with text, image and sound* (pp.131-151). London: Sage.

Boyatzis, R. E. (1998). *Transforming qualitative information*. Sage: London

Recommended websites

www.gsu.edu/~wwwcom/

www.qualitative-research.net/fqs/

References

Bauer, M. W. (2000). Classical content analysis: a review. In M. W. Bauer and G. Gaskell (eds) *Qualitative researching with text, image and sound* (pp.131-151). London: Sage.

Berelson, B. (1952). *Content analysis in communications research*. New York: Free Press.

Boyatzis, R. E. (1998). *Transforming qualitative information*. Sage: London

Dey, I. (1993). *Qualitative data analysis: A user-friendly guide for social scientists*. London: Routledge.

Hollway, W. and Jefferson, T. (2000). *Doing qualitative research differently*. London: Sage.

Huberman, A. M. & Miles, M. B. (1994) *Data management and analysis methods* (pp.428-444). London: Sage.

Joffe, H. (1999). *Risk and 'the Other'*. Cambridge: Cambridge University Press.

Joffe, H & Bettega, N. (Under review). Representations of AIDS among Zambian adolescents. *Journal of Health Psychology*

Joffe, H. & Haarhoff, G. (2002). Representations of far-flung illnesses: the case of Ebola in Britain, *Social Science & Medicine*, Vol 54/6, pp 955-969.

Krippendorff, K. (1980). *Content analysis: an introduction to its methodology*. London: Sage.

Silverman, D. (1993). *Interpreting qualitative data*. London: Sage.

Smith, C.P. (2000). Content analysis and narrative analysis. In H.T.Reis & C.M.Judd (Eds.) *Handbook of research methods in social and personality psychology*. Cambridge: Cambridge University Press.

Glossary terms

Content analysis: a method of data analysis that involves categorising and quantifying the characteristics of qualitative data that are of interest to the researcher

Thematic analysis: a method of qualitative data analysis that involves systematically identifying and describing themes or patterns in a qualitative data set

Code/coding: the procedure of identifying and labelling recurrent features or patterns in a qualitative data set. This allows the researcher to systematically retrieve every instance of that feature or pattern for further analysis (e.g. counting occurrences of that code; comparing instances of that code; carrying out more detailed analysis of the coded segment).

Manifest: a property (e.g. of data) that is obvious and directly observable

Latent: a property (e.g. of data) that is not directly observable but can be inferred

Inductive coding: the procedure of developing coding categories to describe and discriminate between features or patterns that can be discerned in a qualitative data set

Deductive coding: the procedure of applying to a qualitative data set pre-existing coding categories that have been developed prior to data collection on the basis of theory or previous empirical research

Exclusive coding: a coding system in which only one code can be applied to each segment of data

Coding frame: the labels and definitions for the complete set of codes applied to a qualitative data set in a research project

Splitting: the procedure of making distinctions within a set of coded data segments to create two or more subsets labelled with different codes

Splicing: the procedure of merging two or more sets of related coded data segments into a single set labelled with a single code

Inter-rater reliability: the procedure of calculating the correspondence between the codes or ratings independently assigned to the same data by two different people. If both people assign the same codes to the same data segments then the coding or rating system can be considered to be reliable.

Box 1. Key features of coding

Coding involves noting patterns in the data and labelling these patterns to allow distinctions to be drawn and research questions to be answered

The researcher must decide whether to code manifest or latent themes, using deductive or inductive coding categories

As coding progresses, categories are refined by splitting, splicing and linking codes

The codes are described in a coding frame, which should list their labels, detailed definitions, and one or two example text segments

Checking the inter-rater reliability of coding ensures that coding decisions are made explicit and consistent

Box 2. Key features of content analysis and thematic analysis

Content analysis is a method for counting particular features of a text or visual image. Statistical tests can be used to analyse the frequency of codes when content analysis has been carried out on a large, representative data set and the codes have been shown to have good inter-rater reliability.

Thematic analysis is similar to content analysis, but also involves more explicit qualitative analysis of the meaning of the data in context. It is useful for systematically identifying and describing features of qualitative data, which recur across many participants.

When using both methods it is important a) to examine all the data carefully to 'test' how much of it fits the description presented in the analysis, and b) to remember that coding is an interpretive process, and that the frequency of codes does not necessarily reflect their importance.

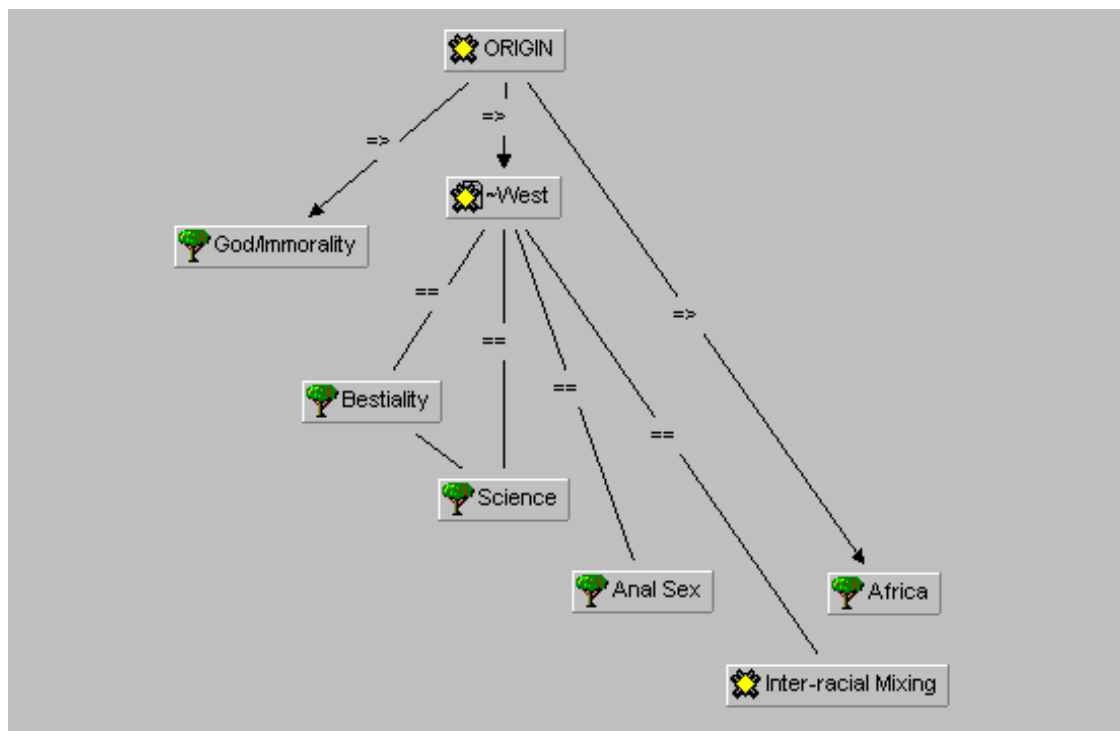
Computer packages to assist qualitative data analysis help the researcher to retrieve relevant text segments for analysis, and to assess the frequency and co-occurrence of codes, but cannot fulfil the central task of qualitative analysis; i.e. interpreting the textual or visual data.

Figure 1: example from a coding frame relating to representations of origin of AIDS

Code Name	Description	Examples
<i>Origin-geographical-Europe</i> <i>Origin-geographical-US</i>	AIDS came from Europe AIDS came from the US	'I only know it came from Europe and not from Africa' 'Though people think that AIDS came from the black man, it actually came from the white men in USA, I think New York'
<i>Origin-geographical-Australia</i> <i>Origin-geographical-Africa</i>	AIDS came from Australia AIDS came from Africa	'It came from Australia' 'monkey was a male and happened to rape a black African woman'
<i>Origin-God/Immorality</i>	AIDS is a punishment/result of the immorality of people. Includes sex before/outside of marriage.	'its God's disease', 'it is God-given, God is the one who has brought this disease as a punishment for those people who like moving up and down'
<i>Origin-practice-bestiality</i>	Includes sex with monkeys, chickens, dogs	'they went and slept with some monkeys and then those monkeys were said to have a certain disease... those diseases were passed on to those people'
<i>Origin-practice-science</i>	AIDS was manufactured in a laboratory by scientists	'AIDS was scientific, I hear some people were carrying out an experiment'
<i>Origin-practice-anal sex</i>	AIDS is a result of anal sex	'I think it came by having anal sex... I think it came from those people like homosexuals'
<i>Origin-practice-mixing</i>	AIDS is a result of interracial sex	'it came from the white people, its like they were mixing with us Africans'

(reproduced from Joffe and Bettega, under review)

Figure 2: The origin of AIDS according to a sample of urban, Zambian adolescents



Key: => *Caused by*
 == *Associated with*

(reproduced from Joffe and Bettega, under review)