

Best practices for analyzing open-ended questions

Your guide to making sense of free-text feedback



Contents

What is coding and why does it matter?	3
Two key approaches to coding open-ended responses	4
3 key rules for coding open-ended questions	5
What is a coding frame?	7
Advantages and disadvantages of code frames	8
Coverage and flexibility of a coding frame	9
Deductive and inductive coding styles	10
4 Ways to choose high quality codes	11
Accuracy of coding open-ended questions	12
6 takeaways to remember	13

What is coding and why does it matter?

When you hear a term like 'big data' it almost always refers to quantitative data: numbers or categories. Quantitative data is easily understood and processed by statistical and machine learning techniques, but it can fail to capture the complete story. Qualitative data like free text responses are harder to process but contain valuable information. Dealing with it is hard but important!

Qualitative researchers believe that numbers won't get you far. They believe that by interviewing people and asking them to answer open-ended questions, you can learn more than by only looking at quantitative data.

Let's take NPS surveys for example. The NPS score, calculated from numeric answers to 'How likely on a scale from 0 to 9 are you to recommend us to friends or family?' will result in a single measure of company's performance. But it's the open-ended answers to the question 'Why did you give us that score?' that will teach you how to improve that measure in the future.

Qualitative research produces a lot of text. Researchers use **coding** to draw conclusions from this data. Survey questions where respondents can write whatever they like are also called open-ended questions. Each of the responses is known as a verbatim. 'Coding' or 'tagging' each response with one or more codes helps capture what the response is about, and in turn, summarise the results of the entire survey effectively. If we compare coding to NLP methods for analyzing text, in some cases coding can be similar to text categorization and in others to keyword extraction. Next, let's look at coding and the different methodologies in more detail.

Two key approaches to coding open-ended responses

Open-ended survey questions often provide the most useful insights, but if you're dealing with hundreds or even thousands of answers, summarizing them will give you the biggest headache. (Trust us, we know!). The answer, then, lies in coding open-ended questions. Easier said than done – where to start?

Whether you go for manual or automated coding, it's a good idea to learn best practices from people who have been dealing with text for decades, **qualitative researchers**. In this e-book, you'll get the low down on how manual coding works.

Manual coding was invented many years ago in qualitative research. Here, a person codes each free-text response with one or more categories. If you are dealing with less than a few hundred responses, you can do it yourself, or you can hire an agency.

Automated coding has become popular in recent years. This is an area of text analytics where algorithms are used to do this task. Data scientists could use a generic NLP library, but tuning such libraries is hard, and the results are often difficult to interpret. Solutions like [Thematic](#) use algorithms designed specifically for coding people's responses. We recently published a [white paper explaining the reliability of automated coding](#) and ways of measuring it objectively.

3 key rules for coding open-ended questions

Whether you go for the manual or automated approach, it's a good idea to learn best practices from people who have been dealing with text for decades. Our main recommendations are:

1 **Fight Bias**

Don't start with a conceived notion of what you are hoping to find. Make sure categories or themes emerge from data rather than pre-defining them.

2 **Provide Coverage**

Make sure that the code frame covers a wide range of contrasting categories. For example, a technology company may have categories related to the UI of the product, customer service, the brand itself, stability, specific features and so on.

3 **Ensure Consistency**

If new categories emerge over time, all previously coded responses need to be re-analyzed. Keep iterating until each customer response is tagged with all themes present in the dataset.

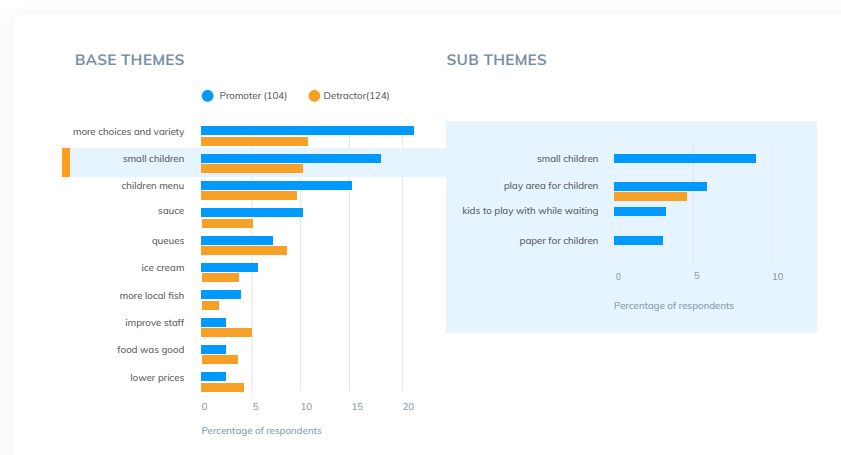
How to use coding results to make sense of customer comments?

Either you will use an existing dashboard, or you create one yourself using Excel, Numbers or Tableau.

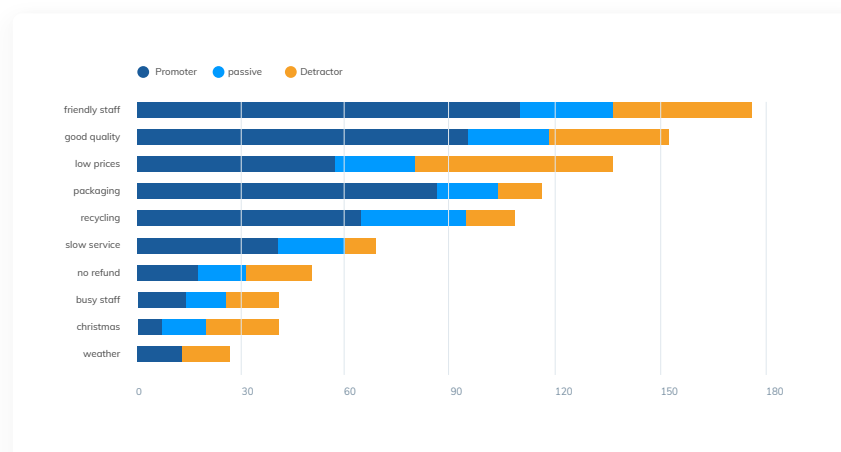
Two key things to capture:

- What is the overall importance of each category in your code frame?
- How do these changes depend on the segments of data?

The best chart to capture this is the old good bar chart. Here is an example of what it could look like in a custom dashboard:



Example of comparing themes of Promoters and Detractors



And here is an example of how it could look in Excel or Numbers:

Comparing theme distribution for Promoters, Passives, and Detractors using Numbers. Of course, what it looks like is not as important as whether the analysis is accurate and useful.

What is a coding frame?

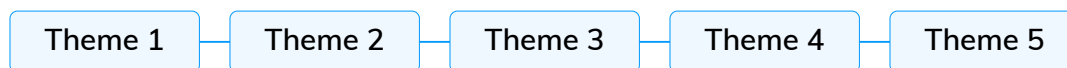
Now, let's look at coding and the different methodologies in more detail. We often refer to how to perform the task manually, but if you are looking at using an automatic solution, this knowledge will help you understand what matters and how to choose an approach that's effective.

When creating codes, they are put into a coding frame. This frame is important because it represents the organizational structure and influences the usefulness of the coded results.

There are two types of frame: 'flat' and 'hierarchical':

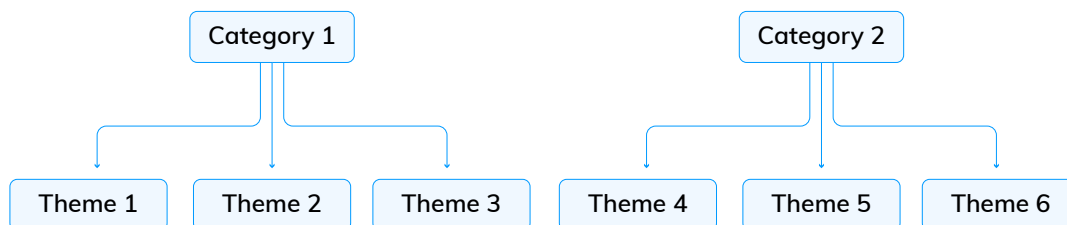
Flat frame

A Flat frame means that all codes are of the same level of specificity and importance. It is easy to understand. But if it gets large, organizing and navigating it will be difficult.



Hierarchical frames

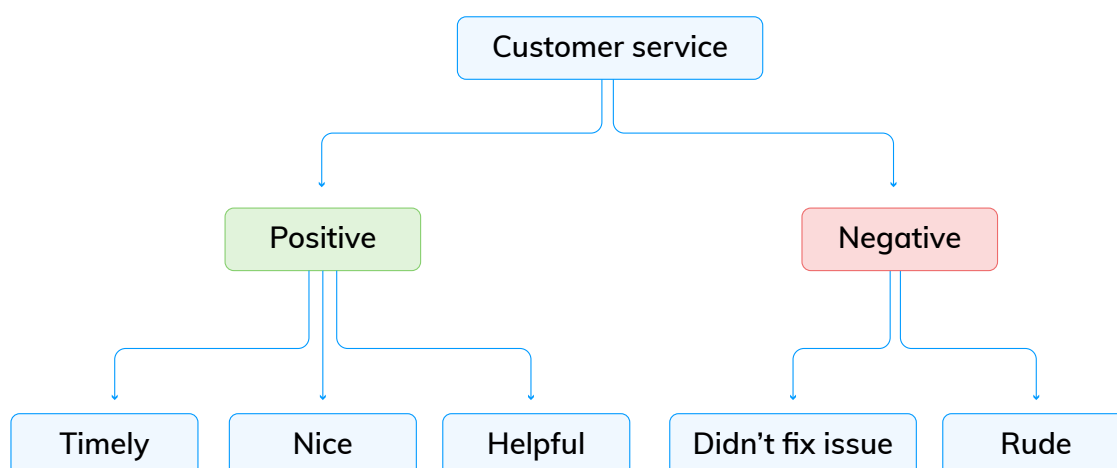
Hierarchical frames capture a taxonomy of how the codes relate to one another. They allow you to apply a different level of granularity during coding and analysis of the results.



Advantages and disadvantages of code frames

Flat code frame	Hierarchical code frame
Supports fewer codes	✓ Supports a larger code frame
✓ Easier and faster to manually code with	Requires navigating the code frame to the right one
✓ Easier to provide consistent coding	Prone to subjective opinion of how each answer is coded
Difficult to capture answers that aren't common leading to large 'other' category	✓ Can organize on basis of organizational structure etc.
Doesn't differentiate between importance and levels of specificity of themes	✓ Allows for different levels of granularity

One interesting application of a hierarchical frame is to support sentiment differences. If the top-level code describes what the response is about, a mid-level one can specify if it is positive or negative and a third level the attribute or specific theme. You can see an example of this type of coding frame below.



Using Sentiment in a Hierarchical coding frame

Coverage and flexibility of a coding frame

One critical thing to consider when coding open-ended questions is the size and the coverage of the frame.

Make sure to group responses with the same themes, disregarding of wording, under the same code. For example, a code 'cleanliness' could cover responses mentioning words like 'clean', 'tidy', 'dirty', 'dusty' and phrases like 'looked like a dump', 'could eat off the floor'. The coder needs a good understanding of each code and its **coverage**. Having few codes and a fixed frame makes the decision easier. Having many codes, particularly in a flat frame, makes it harder as there can be ambiguity and sometimes it isn't clear what exactly a response means. Manual coding also requires the coder to remember or be able to find all the relevant codes which is harder with a large coding frame.

Finally, coding frames should be **flexible**. Coding a survey is a costly task, especially if done manually, and so the results should be usable in different contexts. Imagine this: You are trying to answer the question 'what do people think about customer service' and create codes capturing key answers. Then you find that the same survey responses also have many comments about your company's products. If you need to answer 'what do people say about our products', you may find yourself having to code from scratch! Creating a coding frame that is flexible and has good coverage (see the Inductive Style section below) is a good way to get value in the future.

Deductive and inductive coding styles

What are the two approaches to manual coding open-ended questions, and which one is best?

Deductive coding using pre-existing frame

With deductive coding, you start with a predefined set of codes. These might come from an existing taxonomy that may cover departments in a business or industry-specific terms. Here, codes are driven by a project objective and are intended to report back on specific questions. For example, if the survey is about customer experience and you already know that you are interested in problems that arise from call wait times then this would be one of the codes.

The deductive approach has the benefit that you can guarantee the items you are interested in will be covered, but you need to be careful of bias. When you use a pre-existing coding frame, you are starting with a bias as to what the answers could be and might miss themes that would emerge naturally from people's responses.



When you use a pre-existing coding frame, you are starting with a bias as to what the answer could be and might miss theme that would emerge naturally from people's responses

Inductive coding using sampling and re-coding

The alternative coding style is inductive, which is often called 'grounded'. Here, you start from scratch, and all codes arise directly from the survey responses. The process for this is iterative:

1. You read a sample of the data
2. Create codes that will cover the sample
3. Reread the sample and apply the codes
4. Read a new sample of data applying the codes and noting where codes didn't match
5. Create new codes
6. Go back and recode ALL responses again
7. Repeat from step 4.

If you happen to add a new code, split an existing code into two, or change its description, make sure to review how this change will affect all responses. Otherwise, the same response near the beginning and the end of the survey could end up with different codes!

4 Ways to choose high quality codes

When deciding what codes to create several things should be considered.

1

Ensure Coverage.

Codes should cover as many survey responses as relevant. The code should be more generic than the comment itself to allow it to cover other responses. Of course, this needs to be balanced with the usefulness for analysis. For example 'Product' is a very broad code that will have high coverage but limited value. On the other hand 'Product stops working after using it for 3 hours' is very specific and is unlikely to cover many responses.

2

Avoid Commonality.

Having similar codes is ok. But make sure there is a clear difference between them. In maths, this is referred to as orthogonality and captures how independent two things are. 'Customer Service' and 'Product' would be orthogonal while 'Customer service' and 'Customer support' may have subtle differences but are not orthogonal and may work better as the same code.

3

Create contrast.

Try to create codes that contrast with each other. Capture both the positive and negative elements of the same thing separately. For example 'Useful product features' and 'Unnecessary product features' would have contrast.

4

Reduce data.

Let's look at the two extremes: There are as many codes as comments, or each code applies to all responses. In both cases, the coding exercise is pointless. So, try to think about how to reduce the number of data points so that analysis is useful. For example, 'Product stops working after using it for 3 hours' would create an unnecessary data point. Use instead 'Product stops after use'.

Accuracy of coding open-ended questions

Ensuring consistency is hard regardless of whether coding is deductive or inductive. A coder's mental frame and past experiences color how they interpret things. As a result, different people given the same task are very likely to disagree on what the proper codes should be. In fact, one study has shown that the same person coding the same survey on a different day will produce different results.

Mitigate this by logging all decisions and thoughts that went into coding. Review them when applying existing codes or deciding if a new one is necessary. This process will also mean that the choice of codes can be backed up with evidence.

A different, more expensive, approach to ensure accuracy is through careful testing of the coding reliability. The 'test-retest' method involves the same person coding the data twice without looking at the results. The 'independent-coder' method uses a second coder on the same survey. In both cases, the results are then compared for consistency and amended as needed.

6 takeaways to remember

- 1 Coding is the process of assigning codes to open-ended answers, or other types of text data, after which text can be analyzed just like numerical data.
- 2 Code frames can be flat (easier and faster to use) and hierarchical (more powerful).
- 3 Code frames need to have good coverage and be flexible to allow for a complete and a varied analysis of open-ended answers.
- 4 Inductive coding (without a pre-defined code frame) is more difficult but less prone to bias.
- 5 When creating codes make sure they contrast each other and reduce the data.
- 6 Accuracy means consistent coding which can be achieved by logging and reviewing decisions.

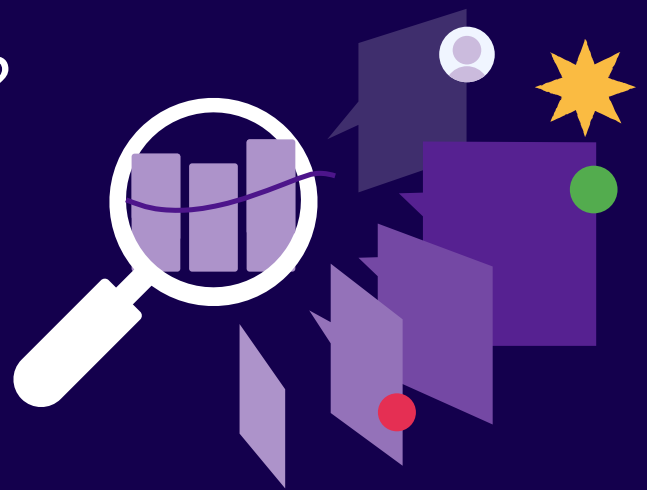


Want to spot customer issues before it's too late?

Thematic allows you to turn customer feedback into actionable insights. We provide a strategic, in-depth analysis of your customer feedback through AI text analytics.

Book a consult with one of our team - we'd be thrilled to show you how Thematic works!

Talk to one of our experts



getthematic.com